IRTG 1792 Discussion Paper 2019-025

SONIC: SOcial Network with Influencers and Communities

Cathy Yi-Hsuan Chen [*]
Wolfgang Karl Härdle [*2] [*3] [*4] [*5]
Yegor Klochkov [*2]

[*] University of Glasgow, UK
[*2] Humboldt-Universität zu Berlin, Germany
[*3] Xiamen University, China
[*4] Singapore Management University, Singapore
[*5] Charles University, Czech Republic

This research was supported by the Deutsche Forschungsgesellschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

SONIC: SOcial Network with Influencers and Communities

Cathy Yi-Hsuan Chen*
University of Glasgow
CathyYi-Hsuan.Chen@glasgow.ac.uk

Wolfgang Karl Härdle†
Humboldt-Universität zu Berlin
haerdle@hu-berlin.de

Yegor Klochkov‡
Humboldt-Universität zu Berlin
klochkoy@hu-berlin.de

September 30, 2019

Abstract

The integration of social media characteristics into an econometric framework requires modeling a high dimensional dynamic network with dimensions of parameter Θ typically much larger than the number of observations. To cope with this problem, we introduce a new structural model — SONIC which assumes that (1) a few influencers drive the network dynamics; (2) the community structure of the network is characterized as the homogeneity of response to the specific influencer, implying their underlying similarity. An estimation procedure is proposed based on a greedy algorithm and LASSO regularization. Through theoretical study and simulations, we show that the matrix parameter can be estimated even when the observed time interval is smaller than the size of the network. Using a novel dataset retrieved from a leading social media platform– StockTwits and quantifying their opinions via natural language processing, we model the opinions network dynamics among a select group of users and further detect the latent communities. With a sparsity regularization, we can identify important nodes in the network.

JEL codes: C1, C22, C51, G41
Keywords: social media, network, community, opinion mining, natural language processing

*Adam Smith Business School, University of Glasgow, UK, corresponding author

†C.A.S.E. - Center for Applied Statistics and Economics, Spandauer Str. 1, 10178 Berlin, Germany; WISE, Xiamen University, China; SKBI, Singapore Management University, Singapore; Faculty of Mathematics and Physics, Charles University, Sokolovska 83, Prague, Czech Republic.

‡C.A.S.E. - Center for Applied Statistics and Economics, Spandauer Str. 1, 10178 Berlin, Germany.
Financial support from the German Research Foundation (DFG) via the International Research Training Group 1792 "High Dimensional Nonstationary Time Series" in Humboldt-Universität zu Berlin is gratefully acknowledged.

1 Introduction

A network is defined through a set of nodes and edges with a given adjacency structure. In a social, financial, or econometric context, such networks are often dynamic and nodes, such as individuals or firms, are changing their activities over time. An analysis of such network dynamics is often based on *vector autoregression*. Consider a network that produces a time series $Y_t \in \mathbb{R}^N$, $t = 1, \dots, T$ and dependencies between its elements are modeled through the equation

$$Y_t = \Theta Y_{t-1} + W_t, \tag{1.1}$$

where W_t are innovations that satisfy $\mathsf{E}[W_t | \mathcal{F}_{t-1}] = 0$, $\mathcal{F}_t = \sigma\{Y_{t-1}, Y_{t-2}, \dots\}$, so that the interactions between the nodes are described by an autoregression operator $\Theta \in \mathbb{R}^{N \times N}$. In terms of the network connections we say that a node i is connected to the node j if

$$\Theta_{ij} \neq 0,$$

so that the nonzero coefficients represent the adjacency matrix of such network, and the sparsity of Θ represents the number of edges. For large-scale time series, one encounters the curse of dimension, as estimating the matrix-parameter Θ with N^2 elements requires a significantly large number of observations T.

Several attempts to reduce the dimensionality have been made in the past literature. Assuming that the elements of a time series form a connected network, Zhu et al. (2017) introduce a Network Autoregression (NAR) with $\Theta_{ij} = \beta A_{ij} / \sum_{k=1}^{N} A_{ik}$, provided that the adjacency matrix $A \in \mathbb{R}^{N \times N}$ is known. Here, the regression operator, defined up to a single parameter β, which is called the *network effect*, can be estimated through simple least squares. Zhu et al. (2019) also extend this model for conditional quantiles. Furthermore, Zhu and Pan (2018) argue that a single network parameter may not be satisfactory as it treats all nodes of the network homogeneously. In particular, the NAR implies that each node is affected by its neighbors in the same extent, while in reality, we may have, e.g., financial institutions that are affected less or more than the others (see Mihoci et al. (2019)). Hence they propose to detect communities in a network based on the given adjacency matrix and suggest that the nodes in each community share a separate network effect parameter. Gudmundsson and Brownlees (2018) take a somewhat opposite direction: their BlockBuster algorithm determines the communities through the estimated autoregressive model, which, however, does not solve the dimensionality problem. Apart from this line of work, sparse regularisations have been extensively used, see Fan et al. (2009); Han et al. (2015); Melnyk and Banerjee (2016).

To sum up, we point out the following problems that one may encounter while dealing with vector autoregression:

- The VAR parameter dimension is significant; one requires even larger time intervals for consistent estimation. Even if one can afford such a dataset, in the long run, autoregressive parametric models tend to be violated, see e.g., Čížek et al. (2009). We, therefore, impose some structural assumptions on the operator Θ, so that estimation through moderate sample sizes is possible.

- The NAR model assumes that the adjacency matrix is known. In particular, this is justified for social networks with a stable and natural friendship/follower-followee relationship. For a network of financial institutions, there is no explicitly defined adjacency matrix, and one has to heuristically evaluate it using additional infor-

mation (identical shareholders, trading volumes) or through analyzing correlations and lagged cross-correlations between returns or risk profiles, see Diebold and Yılmaz (2014) and Chen et al. (2019b). However, there is no rigorous reason to believe that the operator in (1.1) depends explicitly on such an adjacency matrix, see also Cha et al. (2010).

Our main contribution is to propose a new method for modeling social network dynamics, which is a challenging task in the presence of the curse of dimensionality and the absence of knowledge of adjacency. The proposed SONIC — *SOcial Network analysis with Influencers and Communities* has the following advantages. First, it allows us to identify the hidden figures who mainly drive the opinion generating process on social media. Second, it discovers the hidden community structure. The proposed estimation algorithm uncovers the hidden figures and communities simultaneously until the minimal empirical risk is attained. Third, we discuss the theoretical properties and underpinnings to ensure estimation efficiency. We stress that SONIC does not require the imputation of missing data. We demonstrate the applicability of SONIC on a novel social media dataset.

In more detail, the heuristics about the structural assumptions on SONIC are from a realistic view of known facts. Based on well-known user experience on platforms like facebook, twitter, etc., one can assume that some alpha users have significantly more followers than others. Take, for example, celebrities, athletes, analysts, politicians, or Instagram divas. In a network view, these users are the nodes that have much more influence than the rest of the nodes: these nodes are defined as *influencers*. In the framework of autoregression, a node j is an influencer if there is a significant amount of other nodes i such that $\Theta_{ij} \neq 0$. Assuming that the number of influencers is limited, we fix only a few columns of matrix Θ to be significant. This allows us to concentrate on the connections to the influencers, significantly reducing the number of parameters to be estimated. A similar idea is used in Chen et al. (2018), with a group-LASSO regularisation imposed, yielding a solution with few active columns. Notice, however, that relying on the sparsity alone still requires $T > N$, see e.g., Fan et al. (2009); Chernozhukov et al. (2018).

It is also well-known that social networks have small communities, with the nodes exhibiting higher connection density or similar behavior inside communities. Zhu and Pan (2018) make one step to extend the NAR model into a more realistic set-up by allowing separate parameters for each community instead of a single network effect parameter. In our notation, the conditional mean of the response of the node i satisfies

$$\mathsf{E}[Y_{it} | \mathcal{F}_{t-1}] = \Theta_{i1} Y_{it-1} + \cdots + \Theta_{iN} Y_{Nt-1}.$$

Therefore, the behavior of the node i is characterized by the coefficients $\Theta_{i1}, \ldots, \Theta_{iN}$ i.e., the nodes it depends upon. We assume that the nodes are separated into a few clusters such that the nodes from the same cluster share the same dependency structure, which brings a bigger picture into the view: instead of saying that two nodes from the same cluster are more likely to be connected, we say that they connect to the same influencers.

Our primary focus is an application to the opinion dynamics extracted from a microblogging platform dedicated to stock trading, StockTwits (available at `https://stocktwits.com`.) For each user, one can extract the average sentiment score over the messages he posts during the day. Analyzing the resulting high-dimensional time series, on the one hand, we can identify influencers — the users whose opinions are overwhelm-

ingly important, and on the other hand, we determine the community structure. One challenge emerges here: the presence of missing observations since sometimes users do not leave any message. We treat this as follows: assume there is an underlying opinion process that follows network dynamics (1.1). However, such an opinion process might be partially observed, given the random arrival of messages from each user, which results in a commonly used model for missing observations that involve masked Bernoulli random variables. The proposed SONIC accommodates this situation, which seems to have partaken in nowadays social media. We return to it in detail in Section 3.3.

The rest of the paper is organized as follows. Section 2 introduces the readers to the StockTwits platform, describes in detail the available dataset and the process of users' sentiment scores extraction. In Section 3, we first introduce our SONIC model, then describe the estimation procedure and provide a consistency result. In Section 4, we provide simulation results that support the theoretical properties of our estimator. Next, in Section 5, we present and discuss the results of the application of our model to the datasets retrieved from StockTwits. Section 6 concludes. We dedicate Section 7 to the proofs, as well as Sections A, B in the appendix. The reader can find all numerical examples and the SONIC quantlets on `www.quantlet.de`.

2 StockTwits

Social media are an ideal platform where users can easily communicate with each other, exchange information, and share opinions. The increasing popularity of social media is clear evidence of such demand for exchanging opinions and information among granular users in a cyber world. Among social media platforms, we are particularly interested in StockTwits for several reasons. Firstly, it becomes predominantly popular and stands for a leading social network for investors and traders. Secondly, it is similar to Twitter but dedicated to financial discussion. One of the features that lead to its popularity is a well-designed reference between the message content and the referring stock symbols. Conversations are organized around 'cashtags' (e.g., '$AAPL' for APPLE; '$BTC.X' for BITCOIN) that allow to narrow down streams on specific assets. Thirdly and most importantly, users can also express their sentiment/opinions by labeling their messages as 'Bearish' (negative) or 'Bullish' (positive) *via* a toggle button. These are so-called *self-report sentiments*. The available labeled data constitutes an advance on textual analysis that typically relies on the available training dataset. We use this convention and the StockTwits Application Programming Interface (API) to retrieve all messages containing the preferred cashtags. StockTwits API also provides for each message its unique user identifier, the time it was posted within one-second precision and the sentiment associated by the user ('Bullish,' 'Bearish,' or unclassified).

Among over thousand tickers/symbols, we particularly pick up two symbols, $AAPL for APPLE; $BTC.X for BITCOIN, which represents the most popular security and cryptocurrency, respectively. Because they attract investors/users who may possess distinct risk preference, we conjecture that the resulting opinion network and its dynamics may exhibit different structures. In Table 1, we summarize the messages' statistics concerning AAPL and BTC. Even though we exclusively consider these two symbols, the message volume and number of users associated with these two symbols are tremendous. A glimpse of Table 1 reveals different profiles between two symbols. Firstly, the users interested in BTC tend to disclose their sentiment, evident by 44% of labeled messages, while in AAPL only 28% of messages are labeled. It may lead to better training accuracy

in the case of BTC messages relative to the training model based on AAPL. Secondly, there is a clear imbalance between the numbers of positive and negative messages, showing that online investors tend to be optimistic on average, as previously found by Kim and Kim (2014) and Avery et al. (2016). It seems that the imbalance is more evident in the case of AAPL. Judging by the reported average message volume per day, there is no doubt that AAPL can attract more attention than BTC.

Symbols	AAPL	BTC
message volume	449,761	644,597
number of distinct users	26,521	25,492
number of bullish messages	133,316	196,555
number of bearish messages	48,186	90,677
percentage of bullish messages	20.6%	30.4%
percentage of bearish messages	7.4%	14.0%
percentage of labeled messages	28.0%	44.4%
size of positive training dataset	99,985	147,759
size of negative training dataset	36,100	67,752
message volume per day	730	305
number of positive terms in lexicon	4,000	3,775
number of negative terms in lexicon	4,000	3,759
sample period	2017-05-22	2013-03-21
	2019-01-27	2018-12-27

Table 1: Summary statistics of social media messages

2.1 Quantifying message content

There exist two techniques that can converse text data into a quantitative sentiment variable, namely dictionary-based and machine learning-based analysis. Although a machine learning technique has many advantages compared to a dictionary-based approach, the latter offers better transparency, explication, and less computational burden. Loughran and McDonald (2016) recommend that one should consider sophisticated alternative methods only when they add substantive value beyond more straightforward and more transparent approaches, such as the bag-of-word technique. We, therefore, opt for the lexicon approach in the task of sentiment quantification.

A dictionary, or lexicon, is a list of words labeled as positive, negative, or neutral. Given such a list, the *bag-of-words* approach consists of counting the number of positive and negative words in a document in order to assign it a sentiment value or a tone. For example, a simple dictionary containing only the words 'good' and 'bad' with positive and negative labels, respectively, would classify the sentence 'Bitcoin is a good investment' as positive with a tone +1. As the literature (Loughran and McDonald, 2011; Chen et al., 2019a) points out, the simplicity of the dictionary-based approach guarantees transparency and replicability, on the cons side, it comes with limitations associated with natural language analysis. First, referring to Deng et al. (2017) to the 'context of discourse,' one needs to be aware of the content domain, to which language interpretation is sensitive. For example, Loughran and McDonald (2011) point out that words like 'tax' or 'cost' are classified as negative by Harvard General Inquirer lexicon, whereas

they should be considered neutral in the financial context. Another example is about quantifying sentiment toward cryptocurrency, playing the role of non-standard assets and embracing new technologies as part of their characteristics. Chen et al. (2019a) point out that in many domain-specific terms, such as 'blockchain,' 'ICO,' 'hackers,' 'wallet,' 'shitcoin' and 'binance,' 'hodl,' are not covered in the existing financial and psychological dictionaries. They construct a new cryptocurrency lexicon in response to the need of adopting a specific approach to measure sentiment about cryptocurrencies. The second limitation is about the language domain, which Deng et al. (2017) define as the 'lexical and syntactical choices of language.' One example would be the difference between newspapers where one mostly finds a formal and standardized tone, and social media, where slang and emojis prevail (Loughran and McDonald, 2016). As shown by Chen et al. (2019a), online investors also use new 'emojis' such as 🚀 (positive) and 💩 (negative) when talking about cryptocurrencies. These are missing in the traditional dictionary.

To balance the complexity and transparency and also to take into account the domain-specific terms in social media while applying a lexicon approach, in the sentiment quantification for the messages of AAPL we employ the social media lexicon developed by Renault (2017) while in the quantification of BTC messages we advocate the lexicon tailored for cryptocurrency asset by Chen et al. (2019a). Renault (2017) demonstrates that his constructed lexicon significantly outperforms the benchmark dictionaries (Loughran and McDonald, 2016) used in the literature while remaining competitive with more high-level machine learning algorithms. Based on 125,000 bullish and another 125,000 bearish messages published on StockTwits, using the lexicon for social media achieves 90% of classified messages and 75.24% of correct classifications.[1] With a collection of 1,533,975 messages from 38,812 distinct users, posted between March 2013 and December 2018, and related to 465 cryptocurrencies listed in StockTwits[2], Chen et al. (2019a) documents that implementing the crypto lexicon can classify 83% of messages, with 86% of them correctly classified.

The natural language processing (NLP) is a prerequisite for implementing the textual analysis. Following Sprenger et al. (2014) and Renault (2017), we convert unstructured text into clean and manageable textual content as the grounding base throughout the textual analysis. First, all messages are lowercased. To account for lengthening of words, which has been shown to be a critical feature of sentiment expression on microblogs (Brody and Diakopoulos, 2011), but avoid noise in the lexicon, we shrink sequences of repeated letters to a maximum length of 3. Tickers ('$BTC.X,' '$LTC.X,' ...), dollar or euro values, hyperlinks, numbers, and mentions of users are respectively replaced by the words 'cashtag,' 'moneytag,' 'linktag,' 'numbertag,' and 'usertag'. The prefix "negtag_" is added to any word consecutive to 'not,' 'no,' 'none,' 'neither,' 'never,' or 'nobody'. Finally, the three stopwords 'the,' 'a,' 'an' and all punctuation except the characters '?' and '!' are removed. We keep the exclamation and interrogation marks since it has been previously shown that they are often part of significant bigrams that improve lexicon accuracy (Renault, 2017).

The next step is to undertake a lexicon-based approach in order to extract the semantic expression, sentiment, or opinions. For any individual message in Table 1 we filter

[1]The percentage of correct classification is defined as the proportion of correct classifications among all classified messages, while the percentage of classified messages is denoted as the proportion of classified messages among all messages.

[2]This list can be found at https://api.stocktwits.com/symbol-sync/symbols.csv

the terms we collect in the designated lexicon, and equally weight the filtered terms to generate the sentiment score, which also means that the sentiment score of a message is estimated as the average over the weights of the lexicon terms it contains. Recall, that weights of the terms lexicon are in the range of -1 and $+1$. The sentiment score is automatically in the same range.

To visualize the quantified sentiment scores from individuals over time, we select the most active users and display their daily sentiment from 2018-11-01 to 2018-12-27. The heatmap shown in Figure 2.1 is a 2-dimensional matrix with y-axis for user's ID, and x-axis for message posting date, the cell of the heatmap is the quantified sentiment score whose magnitude represented by a color shown in the adjunct color bar. The evolution and dynamics of sentiment among users can be read in such a heatmap presentation. From either Figure 2.1a (AAPL) or Figure 2.1b (BTC), one observes the similar color codes among a subset of users at particular date or period, indicating a contemporaneous common opinion/sentiment and an intertemporal opinion flow among users. Worth noting that some heterogeneity may exist as some users possess optimistic opinions and others are persistently pessimistic.

3 The SONIC model

3.1 Notation

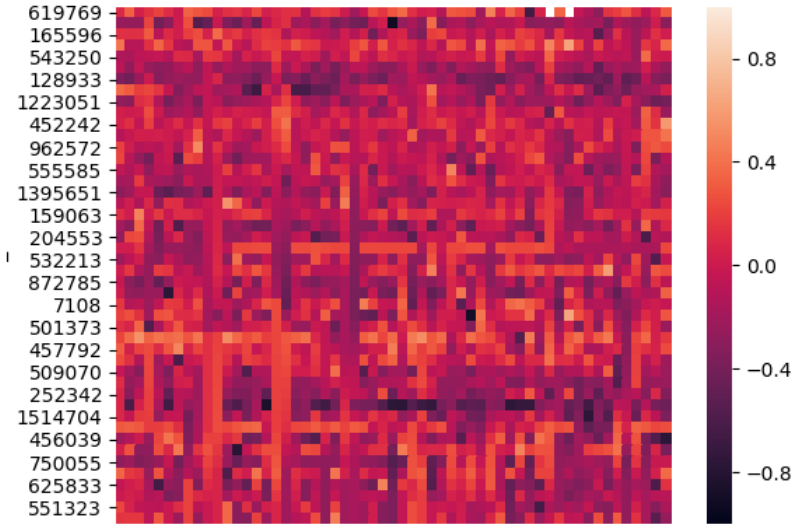
Let us first introduce some basic notations. Through the whole paper, N always denotes the size of the network. Denote by $[N]$ the set of integers from 1 to N, i.e., $[N] = \{1, \dots, N\}$. For a subset of indices $\Lambda \subset [N]$ we denote its complement $\Lambda^c = [N] \setminus \Lambda$. Moreover, if A is a $N \times N$ matrix and $\Lambda_1, \Lambda_2 \subset [N]$ are two subsets of indices, we denote the submatrix $A_{\Lambda_1,\Lambda_2} = (A_{ij})_{i\in\Lambda_1, j\in\Lambda_2}$. We also write for short $A_{\Lambda,\cdot} = A_{\Lambda,[N]}$ and $A_{\cdot,\Lambda} = A_{[N],\Lambda}$.

Furthermore, for a vector $\mathbf{a} \in \mathbb{R}^d$ denote a square matrix $\mathrm{diag}(\mathbf{a}) \in \mathbb{R}^{d\times d}$ that has the values $a_1, \dots, a_d$ on the diagonal and zeros elsewhere. For a square matrix $A \in \mathbb{R}^{d\times d}$ we denote $\mathrm{Diag}(A) \in \mathbb{R}^{d\times d}$ as a diagonal matrix of the same size that coincides with A on the diagonal, i.e., $\mathrm{Diag}(A) = \mathrm{diag}(A_{11}, \dots, A_{dd})$. For the off-diagonal part we use the notation $\mathrm{Off}(A) = A - \mathrm{Diag}(A)$.

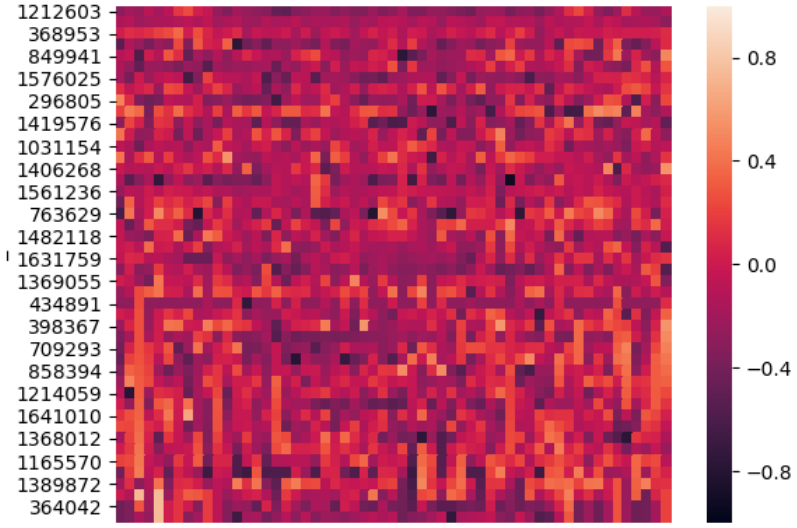
For a real vector $\mathbf{x} \in \mathbb{R}^d$ and $q \geq 1$ or $q = \infty$ denote the ℓ_q-norm $\|\mathbf{x}\|_q = (|x_1|^q + \dots + |x_d|^q)^{1/q}$; for $q = 2$ we ignore the index, i.e., $\|\mathbf{x}\| = \|\mathbf{x}\|_2$; we also denote a pseudo-norm $\|\mathbf{x}\|_0 = \sum_i \mathbf{1}(x_i \neq 0)$. For $A \in \mathbb{R}^{d_1\times d_2}$, $\sigma_1(A) \geq \sigma_2(A) \geq \dots \geq \sigma_{\min(d_1,d_2)}(A)$ denote the non-trivial singular values of A. We will also refer to $\sigma_{\min}(A)$ as the least nontrivial eigenvalue, i.e., $\sigma_{\min}(A) = \sigma_{\min(d_1,d_2)}(A)$. Furthermore, we write $|||A|||_{\mathsf{op}} = \max_j \sigma_j(A)$ for the spectral norm and $|||A|||_{\mathsf{F}} = \mathrm{Tr}^{1/2}(A^\top A) = \left(\sum_{j=1}^{\min(p,q)} \sigma_j(A)^2\right)^{1/2}$ for the Frobenius norm. Additionally, we introduce element-wise norms $\|A\|_{p,q}$ for $p, q \geq 1$ (including ∞) denotes ℓ_q norm of a vector composed of ℓ_p norms of rows of A, i.e., $\|A\|_{p,q} = \left(\sum_i \left(\sum_j |A_{ij}|^p\right)^{q/p}\right)^{1/q}$. Notice that $\|A\|_{2,2} = |||A|||_{\mathsf{F}}$.

3.2 Structural assumptions: Influencers & communities

In our set-up, the behavior of each node $i \in [N]$ is characterized by the coefficients $\Theta_{i1}, \dots, \Theta_{iN}$, and when we group the nodes using their characteristics the notion of



(a) AAPL users



(b) BTC users

Figure 2.1: Social media users' sentiment over time
y-axis is the user's id, while x-axis is time stamp from 2018-11-01 — 2018-12-27.

community is merged with the notion of cluster. We assume that the nodes are separated into clusters, such that these coefficients remain quantitatively comparable for the nodes within each cluster. Let us first give a precise definition of a clustering.

Definition 3.1. *A* K-clustering *of the set of the nodes* $[N]$ *is called a sequence* $\mathcal{C} = (C_1, \dots, C_K)$ *of* K *subsets of* $[N]$*, such that*

- *any two subsets are disjoint* $C_i \cap C_j = \emptyset$ *for* $i \neq j$*;*
- *the union of subsets* C_j *gives all nodes,*

$$C_1 \cup \dots \cup C_K = \{1, \dots, N\}.$$

Two clusterings $\mathcal{C}$ *and* $\mathcal{C}'$ *are equivalent if the corresponding clusters are equal up to a relabelling, i.e., there is a permutation* π *on* $\{1, \dots, K\}$*, such that* $C_j = C'_{\pi(j)}$ *for every* $j = 1, \dots, K$*.*

Furthermore, define a distance between two clusterings as

$$d(\mathcal{C}, \mathcal{C}') = \min_{\pi} \sum_{j=1}^{K} |C_j \setminus C'_{\pi(j)}|.$$

Remark 3.1. The distance between clusterings is, in fact, the minimal amount of node transferring from one cluster to another, that is required to make the clusterings equivalent. To see this, notice that each clustering can be defined as a sequence $(l_1, \dots, l_N)$ of N labels taking values in $\{1, \dots, K\}$, so that each cluster is defined as $C_j = \{i : \ l_i = j\}$. Then, if the clustering $\mathcal{C}'$ corresponds to the labels $l'_1, \dots, l'_N$, the distance between them equals to

$$d(\mathcal{C}, \mathcal{C}') = \min_{\pi} \sum_{i=1}^{N} \mathbf{1}(l_i \neq \pi(l'_i)).$$

We specify our model by imposing structural assumptions concerning the communities and the presence of influencers.

Definition 3.2. *We say that* $\Theta \in \mathsf{SONIC}(s, K)$ *(SOcial Network with Influencers and Communities) if*

- *each user is influenced by at most* s *influencers, i.e.,*

$$\max_i \sum_{j=1}^{N} \mathbf{1}(\Theta_{ij} \neq 0) \leq s;$$

- *there is a* K*-clustering* $\mathcal{C} = (C_1, \dots, C_K)$ *such that*

$$\Theta_{ij} = \Theta_{i'j}, \qquad j = 1, \dots, N$$

whenever i, i' *are from the same cluster* C_l*,* $l = 1, \dots, K$*.*

We will also say that Θ *has clustering* $\mathcal{C}$*.*

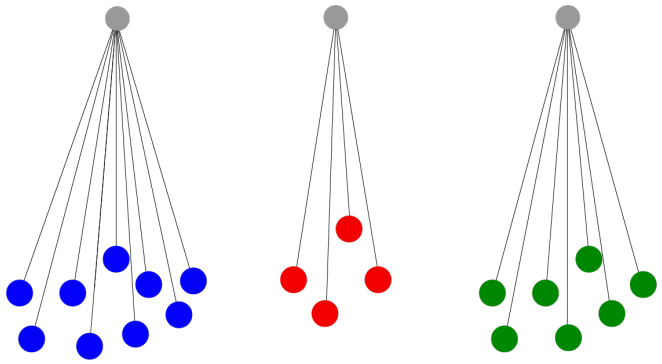


Figure 3.1: Example of a network with influencers for $K = 3$ and $s = 1$.

Once $\Theta \in \mathsf{SONIC}(s, K)$ has clustering $\mathcal{C} = (C_1, \dots, C_K)$, the following factor representation takes place

$$\Theta = Z_{\mathcal{C}} V^\top, \tag{3.1}$$

where $Z_{\mathcal{C}}, V$ are $N \times K$ matrices such that

- $Z_{\mathcal{C}} = [\mathbf{z}_{C_1}, \dots, \mathbf{z}_{C_K}]$ is a normalized index matrix of clustering $\mathcal{C}$, where for any $C \subset [N]$ we denote

$$\mathbf{z}_C = \frac{1}{\sqrt{|C|}} (\mathbf{1}(1 \in C), \dots, \mathbf{1}(N \in C)) \in \mathbb{R}^N$$

  — a normalized index vector for the cluster C and $Z_{\mathcal{C}}^\top Z_{\mathcal{C}} = I_K$;

- $V = [\mathbf{v}_1, \dots, \mathbf{v}_K]$ has sparse columns,

$$\|\mathbf{v}_j\|_0 \le s,$$

  i.e., only a few nodes are active and carrying information;

We present a schematic picture of what we expect in Figure 3.1. Here, the nodes from the same clusters are subject to the same influencers (the grey nodes may be in any of the clusters), which also coincides with the idea of Rohe et al. (2016), who look for the right-hand side singular vectors of the Lagrangian in a directed network, grouping the nodes affected by the same group of nodes.

The equation (3.1) is akin to bilinear factor models, which appear in the econometric literature as a model with factor loadings, see e.g., Moon and Weidner (2018) and the references therein. It is also a popular machine learning technique for low-rank approximation, see a thorough review in Udell et al. (2016). Chen and Schienle (2019) use sparse factors for a closely related model.

3.3 Missing observations

A network of size N represents a multivariate time series $Y_t = (Y_{1t}, \dots, Y_{Nt})^\top \in \mathbb{R}^N$, where Y_{it} is the response of a node $i = 1, \dots, N$ at a time $t = 1, \dots, T$ and contaminated with missing observations. Instead of specifying the exact distribution under the parametric model (1.1), we assume there is the a true parameter $\Theta^* \in \mathbb{R}^{N \times N}$ and some

unknown probability measure P with the expectation E, such that under this measure the time series follows the autoregressive equation

$$Y_t = \Theta^* Y_{t-1} + W_t, \tag{3.2}$$

with $\mathsf{E}[W_t | \mathcal{F}_{t-1}] = 0$ for $\mathcal{F}_{t-1} = \sigma(W_{t-1}, W_{t-2}, \dots)$. For the sake of simplicity, we additionally assume that W_t are independent and have $\mathrm{Var}(W_t) = S$ under P. Once $|||\Theta^*|||_{\mathsf{op}} < 1$ the process exists as a converging series

$$Y_t = \sum_{k \geq 0} (\Theta^*)^k W_{t-k}, \tag{3.3}$$

and the covariance of the process reads as

$$\Sigma = \mathrm{Var}(Y_t) = \sum_{k \geq 0} (\Theta^*)^k S \{(\Theta^*)^k\}^\top.$$

For simplicity, we consider *subgaussian* vectors W_t, as it allows us to have deviation bounds for covariance estimation with exponential probabilities. Recall the following definition, that appears, e.g., in Vershynin (2018).

Definition 3.3. *A random vector $W \in \mathbb{R}^d$ is called L-subgaussian if for every $\mathbf{u} \in \mathbb{R}^d$ it holds*

$$\|\mathbf{u}^\top W\|_{\psi_2} \leq L \|\mathbf{u}^\top X\|_{L_2},$$

where for a random variable $X \in \mathbb{R}$ we denote

$$\|X\|_{\psi_2} = \inf \left\{ C > 0 : \ \mathsf{E} \exp \left\{ \left(\frac{|X|}{C} \right)^2 \right\} \leq 2 \right\},$$
$$\|X\|_{L_2} = \mathsf{E}^{1/2} |X|^2.$$

Implementing SONIC is not impeded by the presence of missing data that appear to be one of the features of social media data. We adopt the framework of Lounici (2014) for vectors with missing observations, assuming that each variable Y_{it} is independent and only partially observed with some probability. Formally speaking, instead of having a realization of the whole vector Y_t, we only observe the masked process Z_t defined as

$$Z_t = (\delta_{1t} Y_{1t}, \dots, \delta_{Nt} Y_{Nt})^\top, \qquad t = 1, \dots, T, \tag{3.4}$$

where $\delta_{it} \sim \mathrm{Be}(p_i)$ are independent Bernoulli random variables for every $i = 1, \dots, N$ and $t = 1, \dots, T$ and some $p_i \in (0, 1]$, which means that each variable Y_{it} is only observed with probability p_i independently from the other variables, with $\delta_{it} = 1$ corresponding to the observed Y_{it} and $\delta_{it} = 0$ to the masked Y_{it}. Obviously, the case $p_i = 1$ for every $i = 1, \dots, N$ corresponds to the process without missing observations. Therefore, the framework constituted by (3.4) serves as a generalization of dynamic network models.

Remark 3.2. In terms of the StockTwits world, we interpret the process Y_t as an unobserved underlying *opinion process*. Such an opinion process quantified from the messages is subject to the random arrival of messages, as users disclose their opinions randomly on social media. Although one may restrict the sample to the case of full observation, the statistical inference may be questionable. Also, discarding nodes with very few missing

observations is a waste of available information. Given the fact that some users are more active than others, we need to account for different probabilities p_i.

Suppose that the probabilities p_i are given (otherwise they can easily be estimated) and set $\mathbf{p} = (p_1, \dots, p_N)^\top$. Following Lounici (2014), we set the observed empirical covariance $\Sigma^* = \frac{1}{T}\sum_{t=1}^T Z_t Z_t^\top$ and consider the following covariance estimator,

$$\hat{\Sigma} = \mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Diag}(\Sigma^*) + \mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Off}(\Sigma^*)\mathrm{diag}\{\mathbf{p}\}^{-1}.$$

It is straightforward to assert an unbiased estimator, i.e.,

$$\mathsf{E}\hat{\Sigma} = \Sigma.$$

The state-of-the-art bound for the error of such covariance estimator is inspired by Klochkov and Zhivotovskiy (2018), Theorem 4.2. In the case of independent vectors Y_t and equal probabilities of observations $p_1 = \dots = p_N = p$ they show that for any $u \geq 1$ with probability at least $1 - e^{-u}$ it holds

$$|||\hat{\Sigma} - \Sigma|||_{\mathsf{op}} \leq C|||\Sigma|||_{\mathsf{op}} \left(\sqrt{\frac{\tilde{\mathbf{r}}(\Sigma)\log\tilde{\mathbf{r}}(\Sigma)}{Tp^2}} \bigvee \sqrt{\frac{u}{Tp^2}} \bigvee \frac{\tilde{\mathbf{r}}(\Sigma)\{\log\tilde{\mathbf{r}}(\Sigma) + u\}\log T}{Tp^2} \right),$$

where $\tilde{\mathbf{r}}(\Sigma) = \frac{\mathrm{Tr}(\Sigma)}{|||\Sigma|||_{\mathsf{op}}}$ denotes the *effective rank* of the covariance Σ. Similarly, the effective rank appears as well in the classic covariance estimation problem (i.e., $p = 1$), see, e.g., Koltchinskii and Lounici (2017) who even provide a matching lower bound. Notice that the effective rank takes values between 1 and the rank of Σ. However, if there is no specific restriction on the spectrum of Σ, the effective rank can grow as large as the full dimension N, which means that the bound above can only guarantee the error of order $\sqrt{\frac{N}{Tp^2}}$, not taking into account the logarithms. On the other hand, one only needs to bound the error within specific low-dimensional subspaces. The following theorem provides such deviation bound for the autoregressive process (3.2), and in its turn accounts for possibly distinct probabilities p_i.

Theorem 3.4. *Assume the vectors W_t are independent L-subgaussian and also*

$$|||\Theta^*|||_{\mathsf{op}} \leq \gamma < 1, \qquad p_i \geq p_{\min} > 0.$$

Let $P, Q \in \mathbb{R}^{N\times N}$ be two arbitrary orthogonal projectors of rank M_1, M_2, respectively. Then, for any $u \geq 1$ it holds with probability at least $1 - e^{-u}$,

$$|||P(\hat{\Sigma} - \Sigma)Q|||_{\mathsf{op}} \leq C|||S|||_{\mathsf{op}} \left(\sqrt{\frac{M_1 \vee M_2(\log N + u)}{Tp_{\min}^2}} \bigvee \frac{\sqrt{M_1 M_2}(\log N + u)\log T}{Tp_{\min}^2} \right),$$

where $C = C(\gamma, L)$ only depends on L and γ.

See proof of this result in Section A.

Additionally, we are interested in estimating lag-1 cross-covariance under the same scenario. Namely, based on the sample $Z_1, \dots, Z_T$ and given the probabilities $p_1, \dots, p_N$ we wish to estimate the matrix $A = \mathsf{E}Y_t Y_{t+1}^\top$. Since $\mathsf{E}[Y_{t+1} | \mathcal{F}_t] = \Theta^* Y_t$ for the linear

process (3.3), the corresponding cross-covariance reads as

$$A = \Sigma(\Theta^*)^\top.$$

Consider the following estimator

$$\hat{A} = \mathrm{diag}\{\mathbf{p}\}^{-1} A^* \mathrm{diag}\{\mathbf{p}\}^{-1},$$

where A^* is the observed empirical cross-covariance

$$A^* = \frac{1}{T-1}\sum_{t=1}^{T-1} Z_t Z_{t+1}^\top.$$

For this estimator, we provide an upper-bound, again with a restriction to some low-dimensional subspaces.

Theorem 3.5. *Let P, Q be two projectors of rank M_1 and M_2, respectively. Assume the independent vectors W_t are L-subgaussian and also*

$$|||\Theta^*|||_{\mathsf{op}} \le \gamma < 1, \qquad p_i \ge p_{\min} > 0.$$

Then, for any $u \ge 1$ it holds with probability at least $1 - e^{-u}$

$$|||P(\hat{A} - A)Q|||_{\mathsf{op}} \le C|||S|||_{\mathsf{op}} \left(\sqrt{\frac{(M_1 \vee M_2)(\log N + u)}{T p_{\min}^2}} \bigvee \frac{\sqrt{M_1 M_2}(\log N + u)\log T}{T p_{\min}^2} \right),$$

where $C = C(\gamma, L)$ only depends on γ and L.

We postpone the proof to Section A.

3.4 Alternating minimization algorithm

In order to estimate the matrix $\Theta = Z_{\mathcal{C}} V^\top$, we need to estimate both $\mathcal{C}$ and V simultaneously. Suppose that we have some clustering $\mathcal{C}$ at hand and we aim to estimate the corresponding V. The mean squared loss from the fully observed sample would look like

$$R^*(V; \mathcal{C}) = \frac{1}{2(T-1)} \sum_{t=1}^{T-1} \|Y_{t+1} - Z_{\mathcal{C}} V^\top Y_t\|^2 \tag{3.5}$$

$$= \frac{1}{2}\mathrm{Tr}(V^\top \tilde{\Sigma} V) - \mathrm{Tr}(V^\top \tilde{A} Z_{\mathcal{C}}) + \frac{1}{2(T-1)} \sum_{t=1}^{T-1} \|Y_{t+1}\|^2, \tag{3.6}$$

where we used the fact that $Z_{\mathcal{C}}^\top Z_{\mathcal{C}} = I_K$ and the trace of a matrix product is invariant with respect to transition $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$. Here, we also denote

$$\tilde{\Sigma} = \frac{1}{T-1}\sum_{t=1}^{T-1} Y_t Y_t^\top, \qquad \tilde{A} = \frac{1}{T-1}\sum_{t=1}^{T-1} Y_t Y_{t+1}^\top,$$

to be empirical covariance and empirical lag-1 covariance built on a sample $Y_1, \dots, Y_T$, respectively, which we observe only partially. In reality, the feasible estimators are $\hat{\Sigma}$ and

$\hat{A}$, which we have introduced in the previous section. A natural solution is to plug-in these estimators into the expression (3.5) instead of the unobserved $\tilde{\Sigma}$ and $\tilde{A}$. The last term $\frac{1}{2(T-1)}\sum_{t=1}^{T-1}\|Y_{t+1}\|^2$ does not depend on the parameters $\mathcal{C}$ and V at all; therefore, we can drop it. We end up with the following risk function that we need to minimize,

$$R(V;\mathcal{C}) = \frac{1}{2}\mathrm{Tr}(V^\top \hat{\Sigma} V) - \mathrm{Tr}(V^\top \hat{A} Z_{\mathcal{C}}).$$

In particular, it is not hard to derive from Theorems 3.4 and 3.5 that for any fixed pair $\mathcal{C}, V$ the values of $R(V;C)$ and $R^*(V;\mathcal{C}) - \frac{1}{2(T-1)}\sum_{t=1}^{T-1}\|Y_{t+1}\|^2$ are close with high probability.

As we are searching for a sparse matrix V, we additionally impose a LASSO regularization and end up with the following convex optimization,

$$\begin{aligned}\hat{V}_{\mathcal{C},\lambda} = \arg\min R_\lambda(V;\mathcal{C}), \qquad R_\lambda(V;\mathcal{C}) =& R(V;\mathcal{C}) + \lambda\|V\|_{1,1} \\ =& \frac{1}{2}\mathrm{Tr}(V^\top \hat{\Sigma} V) - \mathrm{Tr}(V^\top \hat{A} Z_{\mathcal{C}}) + \lambda\|V\|_{1,1},\end{aligned}$$

where $\|V\|_{1,1} = \sum_{ij}|V_{ij}|$, and tuning parameter $\lambda > 0$ depends on the dimension N and number of observations T. Concerning this minimization problem, we have the following observations:

- the problem reduces to simple quadratic programming and therefore can be efficiently solved;
- since $\|V\|_{1,1} = \sum_{j=1}^{K}\|\mathbf{v}_j\|_1$ we can rewrite

$$\begin{aligned}R_\lambda(V;\mathcal{C}) =& \frac{1}{2}\mathrm{Tr}\left(V^\top \hat{\Sigma} V\right) - \mathrm{Tr}\left(V^\top \hat{A} Z\right) + \lambda\|V\|_{1,1} \\ =& \sum_{j=1}^{K}\frac{1}{2}\mathbf{v}_j^\top \hat{\Sigma}\mathbf{v}_j - \mathbf{v}_j^\top \hat{A}\mathbf{z}_j + \lambda\|\mathbf{v}_j\|_1.\end{aligned}$$

  Therefore, we need to solve K independent problems of size N, which reduces computational complexity, and one may also be implement it in parallel.

Ideally, we want to solve the following problem (note that the number of clusters K and the tuning parameter λ are fixed)

$$F_\lambda(\mathcal{C}) \to \min_{\mathcal{C}}, \qquad F_\lambda(\mathcal{C}) = \min_{V} R_\lambda(V;\mathcal{C}).$$

We can employ a simple greedy procedure. In the beginning, we initialize $\mathcal{C}^{(0)} = (l_1,\dots,l_N)$ randomly; each label takes values $1,\dots,K$. Then, at a step t, we try to change one label of a node that reduces the risk the most, in other words, we try all the clusterings in the nearest vicinity of the current solution $\mathcal{C}^{(t)}$, i.e.,

$$\mathcal{C}^{(t+1)} = \arg\min_{d(\mathcal{C},\mathcal{C}^{(t)})\le 1} F_\lambda(\mathcal{C}).$$

At each such step, we would need to calculate $F_\lambda(\mathcal{C})$ for $\mathcal{O}\{N(K-1)\}$ different candidates.

Remark 3.3. In general, it is impossible to optimize arbitrary function $f(\mathcal{C})$ with respect

to a clustering. The K-means is well-known to be NP-hard, however, different solutions are widely used in practice, see Shindler et al. (2011) and Likas et al. (2003).

To speed up the trials of the greedy procedure, we utilize an alternating minimization strategy. Suppose, in the beginning, we initialize the clustering by $\mathcal{C}^{(0)}$ and compute the LASSO solution $V^{(0)} = V_{\mathcal{C}^{(0)},t}$. When updating the clustering, we fix the matrix $V = V^{(t)}$ and solve the problem

$$R_\lambda(V;\mathcal{C}) = \frac{1}{2}\mathrm{Tr}(V^\top \hat{\Sigma} V) - \mathrm{Tr}(V^\top \hat{A} Z_{\mathcal{C}}) + \lambda \|V\|_{1,1} \to \min_{\mathcal{C}},$$

where only the term $-\mathrm{Tr}(V^\top \hat{A} Z_{\mathcal{C}})$ depends on $\mathcal{C}$. Minimizing by conducting a few steps of the greedy procedure we obtain the next clustering update $\mathcal{C}^{(t+1)}$. Then, we again update the V-factor by setting $V^{(t+1)} = V_{\mathcal{C}^{(t+1)},\lambda}$. We continue so until the clustering does not change or the number of iterations exceeds a specific limit. The pseudo-code in Algorithm 1 summarizes this procedure.

Result: a pair $(\hat{\mathcal{C}}, \hat{V})$
initialize $\mathcal{C}^{(0)} = (l_1^{(0)}, \dots, l_N^{(0)})$ randomly;
$t \leftarrow 0$;
while $t < max\_iter$ **do**
  update $\hat{V}^{(t)} \leftarrow \arg\min R_{\mathcal{C}^{(t)},\lambda}(V)$;
  for $i = 1, \dots, N$ **do**
    for $l = 1, \dots, N$ **do**
      consider candidate $\mathcal{C}' = (l_1^{(t)}, \dots, l_{i-1}^{(t)}, l, l_{i+1}^{(t)}, \dots, l_N^{(t)})$;
      $r_{il} \leftarrow -\mathrm{Tr}(V^{(t)} \hat{A} Z_{\mathcal{C}'})$;
    end
  end
  $(i^*, l^*) = \arg\min r_{il}$;
  update $\mathcal{C}^{(t+1)} \leftarrow (l_1^{(t)}, \dots, l_{i^*-1}^{(t)}, l^*, l_{i^*+1}^{(t)}, \dots, l_N^{(t)})$;
  if $\mathcal{C}^{(t+1)} = \mathcal{C}^{(t)}$ **then**
    return $(\mathcal{C}^{(t)}, V^{(t)})$;
  else
    $t \leftarrow t + 1$;
  end
end

Algorithm 1: Alternating greedy clustering procedure.

3.5 Local consistency result

In this section, we show the existence of a locally optimal solution in the neighborhood of the true parameter with high probability. We call a clustering solution $\hat{\mathcal{C}}$ *locally optimal* if the functional $F_\lambda(\cdot)$ has the minimum value at point $\hat{\mathcal{C}}$ among its nearest neighbours $d(\mathcal{C}, \hat{\mathcal{C}}) \le 1$. In particular, Algorithm 1 stops at such a solution.

Conditions

Here we describe the conditions that we need for the consistency result. The first condition concludes the requirements of Theorems 3.4 and 3.5.

Assumption 1. *There is some $\Theta^* \in \mathbb{R}^{N\times N}$ such that $|||\Theta^*|||_{\mathsf{op}} \le \gamma$ for some $\gamma < 1$ and the time series Y_t follows (3.3). The innovations W_t are independent with $\mathsf{E}W_t = 0$ and $Var(W_t) = S$. Moreover, each W_t is L-subgaussian.*

Furthermore, we impose structural assumptions on the true parameter Θ^* described in Section 3.2.

Assumption 2. *The true VAR operator admits decomposition with K-clustering $\mathcal{C}^*$*

$$\Theta^* = Z_{\mathcal{C}^*} V^*,$$

and meets the following conditions:

1. $|||\Theta^*|||_{\mathsf{op}} = |||V^*|||_{\mathsf{op}} \le \gamma < 1$ *for some constant* $\gamma \in (0,1)$*;*
2. *cluster separation*
$$\sigma_{\min}([V^*]^\top \Sigma V^*) \ge a_0 \tag{3.7}$$
for some $a_0 > 0$;
3. *sparsity: for every $j = 1,\dots,K$ the active set $\Lambda_j = \mathrm{supp}(\mathbf{v}_j^*)$ satisfies*
$$|\Lambda_j| \le s;$$
4. *significant active coefficients: there is $\tau_0 > 0$ such that*
$$|v_{ij}^*| \ge \tau_0 s^{-1/2}, \qquad i \in \Lambda_j, \qquad j = 1,\dots,K\,. \tag{3.8}$$
Here each $\|\mathbf{v}_j^*\| \le 1$ has at most s nonzero values, hence the normalization;
5. *significant cluster sizes: for some $\alpha \in (0,1)$ it holds*
$$\frac{\min_j |C_j^*|}{\max_j |C_j^*|} \ge \alpha.$$

Notice that the condition (3.7) requires that the clusters are appropriately separated, i.e., each $\mathbf{v}_j^*$ is far enough from a linear combination of the rest. Another assumption is concerned with the population covariance Σ.

Assumption 3. *The covariance of Y_t reads as*

$$\Sigma = \sum_{k=0}^{\infty} (\Theta^*)^k S [(\Theta^*)^k]^\top,$$

where $S = Var(W_t)$. We impose the following assumptions on this matrix.

1. *bounded operator norm*
$$|||\Sigma|||_{\mathsf{op}} \le \sigma_{\max};$$
2. *restricted least eigenvalue*
$$\sigma_{\min}(\Sigma_{\Lambda_j,\Lambda_j}) \ge \sigma_{\min}, \qquad j = 1,\dots,K\,.$$

Note that we do not assume that the smallest eigenvalue of Σ is bounded away from zero, but only those corresponding to the small subsets of indices are. For the sake of simplicity, we additionally assume that the ratio

$$\frac{\sigma_{\max}}{\sigma_{\min}} \leq \kappa,$$

is bounded by some constant $\kappa \geq 1$. Additionally, we can treat the values L, γ, a_0, τ_0, and α as constants. Below we focus on to what extent the relationship between N, T, s, K, and the probabilities of the observations p_i, $i = 1, \ldots, N$ allows consistent estimation of the parameter Θ.

Finally, we present the assumption that allows controlling the exact recovery of sparsity patterns for the LASSO estimator.

Assumption 4. *For every $j = 1, \ldots, K$ it holds*

$$\|\Sigma_{\Lambda_j^c, \Lambda_j} \Sigma^{-1}_{\Lambda_j, \Lambda_j}\|_{1,\infty} \leq \frac{1}{4}.$$

Recall that Λ^c is the complement of $\Lambda \subset [N]$ in $[N]$.

Remark 3.4. The inequality $\|\Sigma_{\Lambda_j^c, \Lambda_j} \Sigma^{-1}_{\Lambda_j, \Lambda_j}\|_{1,\infty} < 1$ allows us to derive the exact recovery of the sparsity pattern at the LASSO procedure-step described above. In Section B we show a straightforward extension of the results from Tropp (2006) to the case with the presence of missing observations.

Theorem 3.6. *Suppose that Assumptions 1-4 hold. There are constants $c, C > 0$ that depend on L, γ such that the following holds. Suppose,*

$$\sqrt{\frac{sn^* \log N}{T p_{\min}^2}} \bigvee \sqrt{\frac{s \log N \log^2 T}{T p_{\min}^2}} \leq c, \tag{3.9}$$

where $n^* = \max_{j \leq K} |C_j^*|$ and, additionally, $N \geq (C\alpha^2 \vee \kappa)K$. Then, with probability at least $1 - 1/N$ for any λ in the range

$$C\sigma_{\max} \sqrt{\frac{\log N}{T p_{\min}^2}} \leq \lambda \leq c \left\{ \kappa^{-4} (a_0^2 / \sigma_{\max}) K^{-2} s^{-1} \bigwedge \sigma_{\min} \tau_0 s^{-1} \right\}, \tag{3.10}$$

and, additionally, $\lambda \geq C\alpha^2 K/N$, there is a locally optimal solution $\hat{\mathcal{C}}$ satisfying

$$\| \| Z_{\hat{\mathcal{C}}} \hat{V}_{\hat{\mathcal{C}}, \lambda}^\top - \Theta^* \| \|_{\mathsf{F}} \leq \left\{ 3\sigma_{\min}^{-1} \sqrt{Ks} + \frac{C\gamma}{a_0} \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^2 K\sqrt{s} \right\} \lambda \,.$$

Moreover, the exact support recovery takes place, i.e., $\operatorname{supp}(\hat{V}_{\hat{\mathcal{C}}, \lambda}) = \operatorname{supp}(V^*)$.

Let us discuss this result. According to the theorem, a greater λ gives greater error once it is in the required range. This comes naturally, as the result is based on the exact recovery, see e.g., Tropp (2006). Ideally, we want to choose the smallest available value,

$$\lambda^* = C\sigma_{\max} \sqrt{\frac{\log N}{T p_{\min}^2}}. \tag{3.11}$$

In this case, the error of the estimator reads as

$$|||\hat{\Theta}_{\lambda^*} - \Theta^*|||_{\mathsf{F}} \leq C' K \sqrt{\frac{s \log N}{T p_{\min}^2}},$$

where C' does not depend on N, T, K, s. Notice that in the case of precisely known clustering $\mathcal{C}^*$ we only need to estimate the matrix V that consists of at most Ks nonzero parameters. Therefore according to Lemma 7.7, the LASSO estimator must give us

$$|||Z_{\mathcal{C}^*} \hat{V}^\top_{\mathcal{C}^*,\lambda^*} - \Theta^*|||_{\mathsf{F}} = |||\hat{V}_{\mathcal{C}^*,\lambda^*} - V^*|||_{\mathsf{F}} \leq C' \sqrt{\frac{Ks \log N}{T p_{\min}^2}},$$

where we used the fact that $Z_{\mathcal{C}^*}$ has orthonormal columns; see also Melnyk and Banerjee (2016) and Han et al. (2015). We conclude that not knowing the exact clustering provides the estimator that is at most $\sqrt{K}$ times worse.

Let us take a closer look at the condition (3.9). Under the cluster size restriction from Assumption 2, we have that all clusters have the size of order N/K, since

$$\alpha \frac{N}{K} \leq |C_j^*| \leq \alpha^{-1} \frac{N}{K}, \qquad j = 1, \ldots, K.$$

Therefore, say if we ignore missing observations, we only need

$$\frac{(sN/K) \log N}{T} \leq c, \tag{3.12}$$

with some constant c depending on α, enabling the estimation toward the parameters. Therefore, once K is large enough, the estimator works with the corresponding error. Notice that the ℓ_1-regularisation alone requires the number of the observations to be at least the number of edges times $\log N$, see Fan et al. (2009). In our setting, the number of connections is up to Ns, so such condition reads as

$$\sqrt{\frac{sN \log N}{T}} \leq 1.$$

Therefore the SONIC model is an improvement in this regard. Finally, we point out that the conditions of Theorem 3.6 imply some limitations on the size of the network concerning the number of observations. Indeed, using the first part of the condition (3.9) and comparing the lower- and upper-bounds of the condition (3.10), we can easily derive

$$\frac{N^{\frac{4}{5}} s^{\frac{6}{5}} \log N}{T p_{\min}^2} \leq c,$$

where $c > 0$ is a constant that only depends on L, γ, a_0, τ_0, and α. Though we do not state that this condition is necesarry, it is clear that in some cases the estimation is possible even when $N > T$.

4 Simulation study

Take $N = T = 100$ and $s = 1$, while K will vary in a range 2...30. As noted above, e.g., by equation (3.12), a larger amount of the clusters leads to better estimation, and we are particularly interested in capturing this effect through simulations. For every $K = 2, \dots, 30$ we construct the following matrix Θ^*,

- pick clusters C_j^* having approximately the same size $\frac{N}{K} \pm 1$;
- for every $j = 1, \dots, K$ set

$$\mathbf{v}_j^* = 0.5\mathbf{e}_j = (0, \dots, 0.5, \dots, 0)^\top,$$

  with a single nonzero value at the place j, so that $s = 1$.

- by construction we have,

$$|||\Theta^*|||_{\mathsf{op}} = |||V^*|||_{\mathsf{op}} = 0.5, \qquad |||\Theta^*|||_{\mathsf{F}} = |||V^*|||_{\mathsf{F}} = 0.5\sqrt{K}.$$

Furthermore we generate i.i.d. $W_{-19}, W_{-18}, \dots, W_T \sim \mathrm{N}(0, I)$ and set

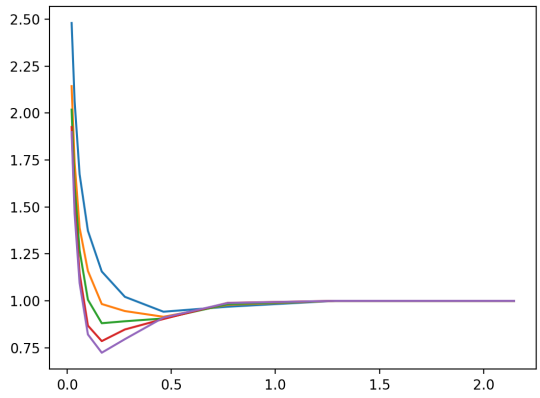
$$Y_t = \sum_{k=0}^{20} (\Theta^*)^k W_{t-k}, \qquad t = 1, \dots T,$$

where due to $0.5^{-20} \approx 10^{-6}$ the terms for $k > 20$ can be neglected. In Figure 4.1a we show the relative error $\mathsf{E}|||\hat{\Theta} - \Theta^*|||_{\mathsf{F}}/|||\Theta^*|||_{\mathsf{F}}$ along regularization paths for different choices of K. Picking the best λ we show the relative error against the number of clusters in Figure 4.1b. We also show that the clustering error $\mathsf{E}d(\hat{\mathcal{C}}, \mathcal{C}^*)$ in Figure 4.1c is subject to the choice of K. All expectations are estimated based on 20 simulations. We conclude that the simulations confirm the following theoretical property of our estimator: the smaller the size of the largest cluster (or equivalent to the case with larger K), the better estimation, while the total size of the network can be even as large as the number of observations.

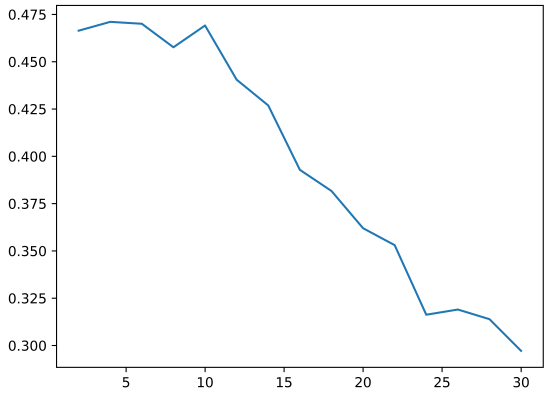
Choice of the regularization parameter λ

It is often suggested to use regularisation $\lambda = \sigma\sqrt{\log N/T}$ in the LASSO literature, where σ stands for a noise level (Belloni and Chernozhukov, 2013; Van de Geer, 2008; Bickel et al., 2009; Van de Geer et al., 2014). In the example above, we have $\sigma = 1$. In our case of missing observations, the value T must be replaced by $Tp_{\min}^2$, which plays the role of "effective" number of observations according to the results of Section 3.3. Furthermore, Wang and Samworth (2018) recommend to disregard multiplicative constants that appear in theory in front of $\sigma\sqrt{\log N/(Tp_{\min}^2)}$ (see equation (3.11)) since it leads to consistent, but rather conservative estimation.

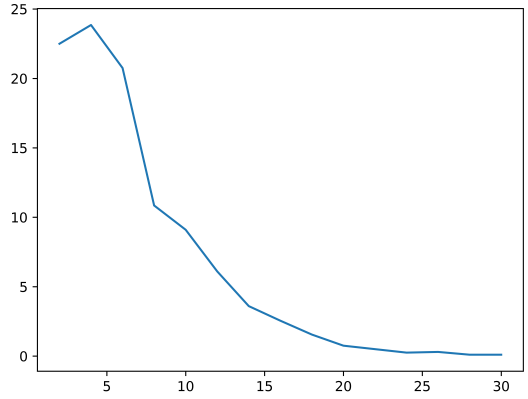
The simulation results support this choice. Let us take a look at the regularisation paths at Figure 4.1a for different values of K. All of the graphs that we show exhibit similar behavior: with λ increasing, the evaluated expected relative loss drops until it reaches its minimum, then it starts to increase until it reaches the constant value that corresponds to $\hat{\Theta}_\lambda = 0$, which obviously happens once the regularization is big enough. Typically, the "oracle" choice corresponds to the minimizer of the expected loss



(a) Expected relative loss $\mathsf{E}\frac{|||\hat{\Theta}-\Theta^*|||_F}{|||\Theta^*|||_F}$ for different λ and $K = 4, 8, 12, 16, 20$.



(b) Expected relative loss $\mathsf{E}\frac{|||\hat{\Theta}_\lambda-\Theta^*|||_F}{|||\Theta^*|||_F}$ for the best λ and $K = 2, \ldots, 30$.



(c) Expected clustering error $\mathsf{E}d(\hat{\mathcal{C}}, \mathcal{C}^*)$ for the best λ and $K = 2, \ldots, 30$.

Figure 4.1: Simulation results for $N = T = 100$ and $s = 1$.

SoNIC_simulation_study

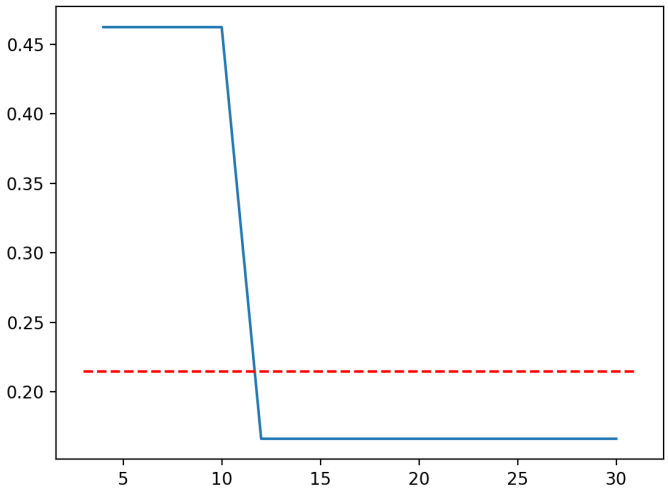


Figure 4.2: The optimal value of λ for $K = 2, \dots, 30$.
SoNIC_simulation_study

$\mathsf{E}|||\hat{\Theta}_\lambda - \Theta^*|||_\mathsf{F}$. In order to compare it with the recommended choice above, we pick for each $K = 2, \dots, 30$ the tuning parameter — among the available choices on the graph — that delivers the minimum to the evaluated expected loss. In Figure 4.2 we show the values of best λ for each $K = 2, \dots, 30$ (blue line) and compare it to the theoretical value $\sqrt{\frac{\log N}{T}}$ (red line). We can see that once the number of clusters is large enough ($K \geq 13$) the corresponding optimal choice of λ approximately equals to $\sqrt{\frac{\log N}{T}}$. On the other hand, as the graph in Figure 4.1c suggests, for $K \leq 12$ the number of nodes assigned to a wrong cluster grows significantly, and one cannot estimate the model with any given regularization parameter.

Remark 4.1. Note that in practice, one must estimate the noise level σ in a data-driven way (Belloni and Chernozhukov, 2013). We suggest to evaluate it using the spectrum of the covariance estimator $\hat{\Sigma}$. One obvious choice can be $\hat{\sigma} = \|\hat{\Sigma}\|$. However, this may lead to an overestimated noise level. We suggest using the following strategy. Since $\Sigma = \Theta^* \Sigma (\Theta^*)^\top + S$, we expect the original covariance to have either K or $K - 1$ spikes (one cluster could be zero). In particular, this is true whenever $S = \sigma I$. We, therefore, suggest using the singular value $\hat{\sigma} = \sigma_K(\hat{\Sigma})$, which means that we avoid the first $K - 1$ components. The resulting regularisation parameter reads as

$$\lambda = \sigma_K(\hat{\Sigma}) \sqrt{\frac{\log N}{T p_{\min}^2}}.$$

In the next section, we stick to this strategy.

5 Application to StockTwits sentiment

Here we present the applicability of SONIC to the dataset described in Section 2. We

concentrate on two network structures comprising the users' tone of messages pointing to AAPL and BTC, respectively. These two symbols, representing the most popular security and cryptocurrency, respectively, may reveal different characteristics that ultimately determine network dynamics. In a sense, two networks may display different communities and influencers, given the symbol's characteristics and the involved users.

Table 1 summarizes the messages' statistics with respect to AAPL and BTC, but further filtering for users and messages ensures the model applicability. Considering missing observations discussed in Section 3.3, we require that the observations are persistent with the same probability p_i over a time period under consideration. Moreover, since in Theorems 3.4 and 3.5 the amount of observations scales with the factor $p_{\min}^2$, we need to avoid the users whose p_i is too small. Based on these remarks, we suggest the following preprocessing steps:

1. pick users with estimated probability $\hat{p}_i \geq 0.5$;
2. select the most extended historical interval over which the user exhibits persistent probability of observation. One can look at a moving average estimation and ensure that for any window it remains within the appropriate confidence interval;
3. take only the users whose historical interval from step 2 is at least 50 days.

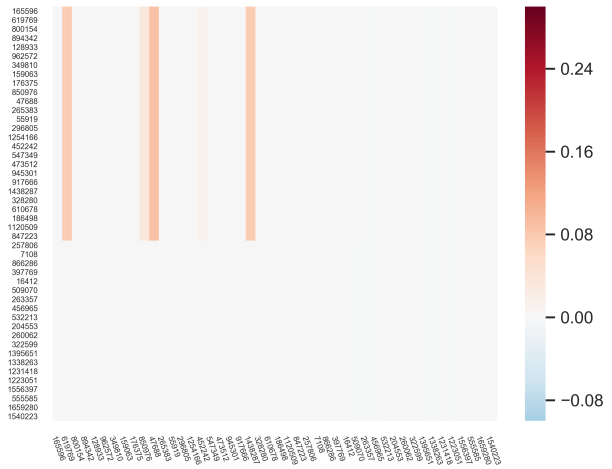
With these criteria, for the AAPL dataset, we are left with 46 users and 72 days, while for BTC, we have 68 users and 52 days. The two datasets are visualized using the heatmap in Figure 2.1.

We apply our SONIC model to AAPL dataset with $\lambda = 0.08$ and $K = 2$. Note that λ is chosen following Remark 4.1, while the stability of the clustering algorithm may guide us the choice of the number of clusters K, which we discuss later. We present a heatmap visualization for the estimated matrix $\hat{\Theta}$ in Figure 5.1a, where we can identify the key users with identification number 47688, 619769, 850976, and 1438287, 547349, 610678[3]. These active nodes are the detected influencers. Looking at their profiles, we end up with the users who have attracted lots of followers, which supports the capability of SONIC. The first four influencers represent trading companies that offer technical and fundamental analysis for the symbols of interest. It shows that investment companies target their potential traders who emerge on the social media platform and influence them strategically. The latter two are individuals who actively post messages. They joined the StockTwits in an early phase, and have successfully collected tremendous "liked" tagged on their messages.

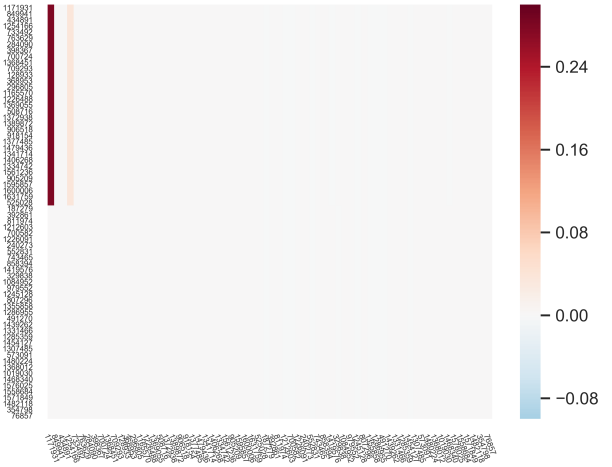
For the BTC dataset, we use $\lambda = 0.1$ and $K = 2$ following the proposed theorems. Figure 5.1b displays the estimated matrix $\hat{\Theta}$ and identifies influencers 1171931 and 1254166. The first one offers the breaking news related to cryptocurrency, while the second one produces the technical analytics to cryptocurrency speculators. Different from what has been identified on AAPL, the influencers on BTC are relatively few but have more dominant power in terms of the magnitude of estimates. Besides, identifying the influencers as individuals seems exceptional.

Interestingly, in both networks we detect two communities, one corresponds to the coefficients $\Theta_{ij} = 0$, simply implying a community comprising of "individualized" users who are neither influencing others nor being influenced by others. They, for instance,

[3]To access a page via identification number type, for example, `https://stocktwits.com/123456` in a browser address line.



(a) AAPL dataset with $N = 46$, $T = 72$ and $K = 2$.



(b) BTC dataset with $N = 68$, $T = 52$ and $K = 2$.

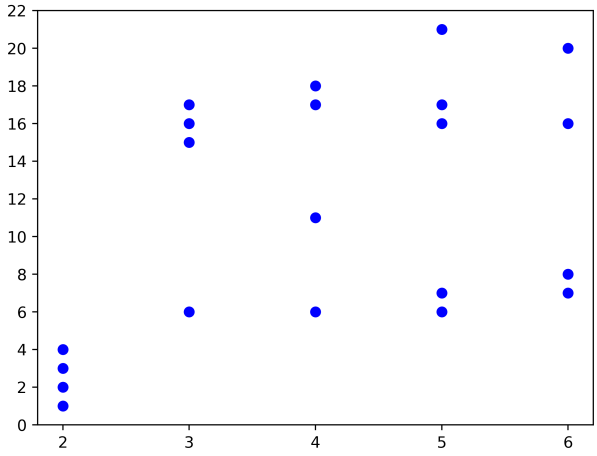
Figure 5.1: Estimated $\hat{\Theta}$ for AAPL and BTC datasets. The axes correspond to user id's and are rearranged with respect to the estimated clusterings.

SoNIC_AAPL_BTC

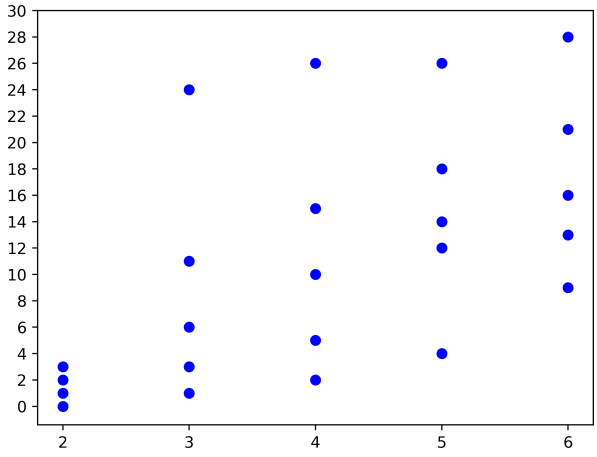
merely react to price changes. On the contrary, the users in the second cluster are more vulnerable to opinion leaders, investment companies spreading the news, and technical reports.

Choosing the number of clusters K

One possible way to decide the number K is to analyze the *stability* of the clustering algorithm (Rakhlin and Caponnetto, 2007; Le Gouic and Paris, 2018). We propose the following procedure. Consider a sequence of intervals $I_1, \dots, I_l \subset \{1, \dots, T\}$ of the same length and let us estimate the clusterings $\hat{\mathcal{C}}_{I_j}$ using the observations $(Y_t)_{t \in I_j}$ for each $j = 1, \dots, l$. If the number of clusters is correct, we expect that the pairwise distances



(a) APPLE dataset



(b) BITCOIN dataset

Figure 5.2: Differences $d(\hat{\mathcal{C}}_{I_1}, \hat{\mathcal{C}}_{I_j})$ for $j = 2, \ldots, 6$ and for $K = 2, 3, 4, 5, 6$.
SoNIC_AAPL_BTC_stability

$\hat{\mathcal{C}}_j$ are small. We take $l = 6$ intervals of length $3T/4 \pm 1$, each of the form

$$I_j = \left[\frac{j}{20} T + 1, \frac{j+14}{20} T \right], \qquad j = 1, \ldots, 6,$$

so that we include all available observations. We then calculate the distances $d(\hat{\mathcal{C}}_{I_1}, \hat{\mathcal{C}}_{I_j})$ for each $j = 2, \ldots, l$ and for different choices of K. The results for both APPLE and BITCOIN are presented in Figure 5.2. The graph demonstrates that the only adequate choice is $K = 2$; for others, the clustering distance reaches the value comparable with the size of the network.

6 Conclusion

Nowadays the interest in dynamics of interaction among the users emerging in social media is dramatically growing as social media becomes an attractive venue where users are allowed to interact with others instantly and intended influencers aggressively show their predominance. The research in this strand is, however, challenging. From an econometric point of view, these dynamics require effective state-of-the-art methodologies that cope with the curse of dimensionality, as well as to characterize psychological interdependence. From a quantitative perspective on, e.g., text-based source channels, like Twitter or StockTwits, the activities, and opinions distilled from social networks boil down to a numerical expression of tone, sentiment or emojis. The joint involvement of these variables constitutes a dynamic network with a possibly growing dimension.

In order to cope with dimensionality in a limited observation setting, we propose SONIC (SOcial Network with Influencers and Communities). SONIC reflects naturally the known facts from social networks: a few influencers and the characterizable homogeneous communities. We provided and discussed several theoretical remarks on the asymptotic consistency of the network parameters. SONIC is based on LASSO regularization with consideration of computational complexity, and we extensively test it in simulations.

Using StockTwits data and their derived sentiments, we deploy a SONIC analysis and display an opinion network for Apple and Bitcoin users (nodes). We detect $K = 2$ communities using stability analysis and discuss the choice of the regularization parameter λ of LASSO.

7 Proof of the main result

This section is devoted to the proof of Theorem 3.6. We start with some preliminary lemmas and then proceed with the proof that consists of several steps. Following the ideas in Gribonval et al. (2015), the proof relies on explicit representation of the loss function.

We exploit the following simplified notation. Denote, $\mathbf{z}_j^* = \mathbf{z}_{C_j^*}$ to be the columns of $Z^* = Z_{\mathcal{C}^*}$ and we also denote $n_j^* = |C_j^*|$ for every $j = 1, \dots, K$. When the clustering $\mathcal{C} = (C_1, \dots, C_K)$ is clear from the context we will also write Z for $Z_{\mathcal{C}}$, $\mathbf{z}_j$ for $\mathbf{z}_{C_j}$, and $n_j = |C_j|$ for every $j = 1, \dots, K$.

7.1 Preliminary lemmas

Lemma 7.1. *Suppose that C_j is such that $\|\mathbf{z}_{C_j} - \mathbf{z}_j^*\| \le 0.3$. Then,*

$$\frac{1}{1.1}|C_j^*| \le |C_j| \le 1.1|C_j^*|.$$

Proof. Suppose, $n_j = |C_j| > n_j^* = |C_j^*|$, then

$$r^2 = \|\mathbf{z}_j - \mathbf{z}_j^*\|^2 = 2 - \frac{2}{\sqrt{n_j n_j^*}}|C_j \cap C_j^*| \ge 2 - 2\sqrt{\frac{n_j^*}{n_j}},$$

since $|C_j \cap C_j^*| \le n_j^*$. Thus, $\sqrt{n_j} - \sqrt{n_j^*} \le (r^2/2)\sqrt{n_j}$, which due to $r \le 0.3$ implies by

rearranging and taking square $n_j \le 1.1 n_j^*$.

If $n_j < n_j^*$ we have,

$$r^2 \ge \|\mathbf{z}_j - \mathbf{z}_j^*\|^2 = 2 - \frac{2|C_j \cap C_j'|}{\sqrt{n_j n_j^*}} \ge 2 - 2\sqrt{\frac{n_j}{n_j^*}},$$

and the fact that $r \le 0.3$ implies $n_j^* \le 1.1 n_j$.

□

Lemma 7.2. *Let* $\|\mathbf{z}_{C_1} - \mathbf{z}_{C_2}\| \le 0.3$. *Then,*

$$\|\mathbf{z}_{C_1} - \mathbf{z}_{C_2}\|_1 \le 1.65\sqrt{N_1}\|\mathbf{z}_{C_1} - \mathbf{z}_{C_2}\|^2 .$$

Proof. Let $N_j = |C_j|$ and $a = |C_1 \cap C_2|$, $b = |C_1 \setminus C_2|$, $c = |C_2 \setminus C_1|$, so that $N_1 = a + b$, $N_2 = a + c$, and $|C_1 \triangle C_2| = b + c$. We have,

$$\|\mathbf{z}_{C_1} - \mathbf{z}_{C_2}\|^2 = \left(\frac{1}{\sqrt{N_1}} - \frac{1}{\sqrt{N_2}}\right)^2 a + \frac{b}{N_1} + \frac{c}{N_2} \ge \frac{b}{N_1} + \frac{c}{N_2} .$$

On the other hand,

$$\begin{aligned}\|\mathbf{z}_{C_1} - \mathbf{z}_{C_2}\|_1 &= \left|\frac{1}{\sqrt{N_1}} - \frac{1}{\sqrt{N_2}}\right| a + \frac{b}{\sqrt{N_1}} + \frac{c}{\sqrt{N_2}} \\ &\le \left|\frac{1}{\sqrt{N_1}} - \frac{1}{\sqrt{N_2}}\right| a + \sqrt{N_1 \vee N_2}\|\mathbf{z}_{C_1} - \mathbf{z}_{C_2}\|^2 .\end{aligned}$$

Since $|N_1 - N_2| \le b + c$ we obviously have,

$$\begin{aligned}\left|\frac{1}{\sqrt{N_1}} - \frac{1}{\sqrt{N_2}}\right| a &= \frac{|N_1 - N_2| a}{\sqrt{(a+b)(a+c)}(\sqrt{a+b} + \sqrt{a+c})} \\ &\le \frac{(b+c)a}{\sqrt{N_1 \vee N_2}\sqrt{a}(2\sqrt{a})} \\ &\le \sqrt{N_1 \vee N_2}\|\mathbf{z}_{C_1} - \mathbf{z}_{C_2}\|^2/2,\end{aligned}$$

and it is left to apply Lemma 7.1. □

Lemma 7.3. *Suppose,* $\frac{\min_j n_j^*}{\max_j n_j^*} \ge \alpha$ *for some* $\alpha \in (0, 1]$ *and let* $\|\mathbf{z}_j - \mathbf{z}_j^*\| \le r$. *Suppose,* $r \le 0.3$. *Then,*

$$\|[Z^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*)\|_1 \le 3.05\alpha^{-1/2} r^2.$$

Proof. 1) We first consider the case $|C_j| = n_j^*$. It holds then

$$[\mathbf{z}_j^*]^\top(\mathbf{z}_j^* - \mathbf{z}_j) = \frac{1}{n_j^*}(n_j^* - |C_j \cap C_j^*|) = \frac{1}{n_j^*}|C_j^* \setminus C_j|.$$

Moreover, for every $k \ne j$ it holds

$$|[\mathbf{z}_k^*]^\top(\mathbf{z}_j^* - \mathbf{z}_j)| = |[\mathbf{z}_k^*]^\top \mathbf{z}_j| = \frac{1}{\sqrt{n_k^* n_j^*}}|C_k^* \cap C_j| \le \frac{\alpha^{-1/2}}{n_j^*}|C_k^* \cap C_j|.$$

Summing up, we get

$$\begin{aligned}\|[Z^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*)\|_1 &\le \frac{\alpha^{-1/2}}{n_j^*}\left(|C_j^* \setminus C_j| + \sum_{k\neq j}|C_k^* \cap C_j|\right)\\ &\le \frac{\alpha^{-1/2}}{n_j^*}\left(|C_j^* \setminus C_j| + |C_j \setminus C_j^*|\right)\\ &= \frac{\alpha^{-1/2}}{n_j^*}|C_j \triangle C_j^*|.\end{aligned}$$

It is left to notice that in the case $|C_j| = |C_j^*| = n_j^*$ we have exactly $\|\mathbf{z}_j - \mathbf{z}_j^*\|^2 = \frac{1}{n_j^*}|C_j \triangle C_j^*|$.

2) Suppose, $n_j = |C_j| > n_j^*$. Obviously, we can decompose $C_j = C_j' \cup B$ such that $|C_j'| = n_j^*$ and $B \cap C_j^* = \emptyset$. Setting $\mathbf{z}_j' = \mathbf{z}_{C_j'}$ we get by the above derivations that $\|[Z^*]^\top(\mathbf{z}_j' - \mathbf{z}_j^*)\|_1 \le \alpha^{-1/2}\|\mathbf{z}_j' - \mathbf{z}_j^*\|^2$. Since $C_j' \cap C_j^* = C_j \cap C_j^*$ we can compare the distances

$$\|\mathbf{z}_j - \mathbf{z}_j^*\|^2 = 2 - \frac{2}{\sqrt{n_j n_j^*}}|C_j \cap C_j^*| > 2 - \frac{2}{n_j^*}|C_j \cap C_j^*| = \|\mathbf{z}_j' - \mathbf{z}_j^*\|^2.$$

Taking the remainder $\mathbf{b} = \mathbf{z}_j - \mathbf{z}_j'$ we have,

$$b_i = \begin{cases} {n_j}^{-1/2} - (n_j^*)^{-1/2}, & i \in C_j',\\ {n_j}^{-1/2}, & i \in B,\\ 0 & \text{otherwise.}\end{cases}$$

Setting $d = n_j - n_j^* = |B|$ it is easy to obtain $|{n_j}^{-1/2} - (n_j^*)^{-1/2}| \le \frac{d}{n_j}\frac{1}{\sqrt{n_j^*}}$. Thus, we get

$$\begin{aligned}\sum_{k=1}^{K}|[\mathbf{z}_k^*]^\top\mathbf{b}| &\le \sum_{i=1}^{k}\frac{1}{\sqrt{n_k^*}}\left(\frac{d}{n_j}\frac{1}{\sqrt{n_j^*}}|C_j' \cap C_k^*| + |B \cap C_k^*|\frac{1}{\sqrt{n_j}}\right)\\ &\le \frac{\alpha^{-1/2}d}{n_j^* n_j}|C_j'| + \frac{\alpha^{-1/2}}{\sqrt{n_j^* n_j}}d\\ &< \frac{2\alpha^{-1/2}d}{\sqrt{n_j n_j^*}}.\end{aligned}$$

We show that the latter is at most $2.05\alpha^{-1/2}r^2$. Indeed, it is not hard to show that from $n_j \le 1.1 n_j^*$ (see Lemma 7.1) it follows

$$\frac{n_j - n_j^*}{\sqrt{n_j n_j^*}} \le 2.05\left(1 - \frac{n_j^*}{\sqrt{n_j n_j^*}}\right) \le 2.05 \times \frac{r^2}{2},$$

thus $\|[Z^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*)\|_1 \le 3.05\alpha^{-1/2}r^2$ and the result follows.

3) The case $n_j < n_j^*$ can be resolved similarly to the previous one. Since $|C_j^* \setminus C_j| \geq n_j^* - n_j$ we can pick a subset $B \subset C_j^* \setminus C_j$ of size $d = n_j^* - n_j$ and set $C_j' = B \cup C_j$ with $|C_j'| = n_j^*$; set also $\mathbf{z}_j' = \mathbf{z}_{C_j'}$. Then, we have

$$\|\mathbf{z}_j' - \mathbf{z}_j^*\|^2 = 2 - 2\frac{|C_j' \cap C_j^*|}{n_j^*} \leq 2 - \frac{2|C_j \cap C_j'|}{\sqrt{n_j n_j^*}} = \|\mathbf{z}_j - \mathbf{z}_j^*\|^2.$$

Thus, by the first part of this proof it holds $\|[Z^*]^\top(\mathbf{z}_j' - \mathbf{z}_j^*)\|_1 \leq \alpha^{-1/2} r^2$. Setting $\mathbf{b} = \mathbf{z}_j' - \mathbf{z}_j$ we have,

$$b_i = \begin{cases} (n_j^*)^{-1/2} - {n_j}^{-1/2}, & i \in C_j, \\ {n_j^*}^{-1/2}, & i \in B, \\ 0 & \text{otherwise.} \end{cases}$$

Since $|{n_j}^{-1/2} - (n_j^*)^{-1/2}| \leq \frac{d}{n_j^*}\frac{1}{\sqrt{n_j}}$ we obtain,

$$\begin{aligned} \sum_{k=1}^{K} |[\mathbf{z}_k^*]^\top \mathbf{b}| &\leq \sum_{i=1}^{k} \frac{1}{\sqrt{n_k^*}} \left(\frac{d}{n_j^*}\frac{1}{\sqrt{n_j}}|C_j \cap C_k^*| + |B \cap C_k^*|\frac{1}{\sqrt{n_j^*}} \right) \\ &\leq \frac{\alpha^{-1/2} d}{(n_j^*)^{3/2} {n_j}^{1/2}}|C_j| + \frac{\alpha^{-1/2}}{n_j^*} d \\ &< \frac{2\alpha^{-1/2} d}{n_j^*}. \end{aligned}$$

It is left to notice that

$$r^2 \geq 2 - \frac{2n_j}{\sqrt{n_j n_j^*}} = \frac{2(\sqrt{n_j^*} - \sqrt{n_j})}{\sqrt{n_j}} = \frac{2(n_j^* - n_j)}{n_j^* + \sqrt{n_j n_j^*}} \geq \frac{2d}{2n_j^*},$$

therefore $\|[Z^*]^\top \mathbf{b}\|_1 \leq 2\alpha^{-1/2} r^2$, thus $\|[Z^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*)\|_1 \leq 3\alpha^{-1/2} r^2$. □

Lemma 7.4. *Let $r = \|\!|Z_{\mathcal{C}} - Z^*|\!\|_{\mathsf{F}}$ and suppose that $r \leq 0.3$. Then $\|\!|P_{\mathcal{C}} - P_{\mathcal{C}^*}|\!\|_{\mathsf{F}}^2 \geq 2r^2(1 - 10\alpha^{-1} r^2)$.*

Proof. Denote $\mathbf{z}_j = \mathbf{z}_{C_j}$ and $r_j = \|\mathbf{z}_j - \mathbf{z}_j^*\|$. It holds,

$$\|\!|P_{\mathcal{C}} - P_{\mathcal{C}^*}|\!\|_{\mathsf{F}}^2 = 2K - 2\mathrm{Tr}(P_{\mathcal{C}} P_{\mathcal{C}^*}) = 2K - \sum_{j,k} (\mathbf{z}_j^\top \mathbf{z}_k^*)^2.$$

Notice, that $2\mathbf{z}_j^\top \mathbf{z}_j^* = 2 - \|\mathbf{z}_j\|^2 - \|\mathbf{z}_j^*\|^2 + 2\mathbf{z}_j^\top \mathbf{z}_j^* = 2 - \|\mathbf{z}_j - \mathbf{z}_j^*\|^2$, i.e., $\mathbf{z}_j^\top \mathbf{z}_j^* = 1 - r_j^2/2$. In particular, $1 - (\mathbf{z}_j^\top \mathbf{z}_j^*)^2 = r_j^2 - r_j^4/4$, whereas $([\mathbf{z}_j^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*))^2 = r_j^4/4$. Since we

additionally have $[\mathbf{z}_k^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*) = [\mathbf{z}_k^*]^\top \mathbf{z}_j$ for $k \neq j$, it holds

$$\begin{aligned}
2K - 2\sum_{j,k}(\mathbf{z}_j^\top \mathbf{z}_k^*)^2 &= 2\sum_j r_j^2 - r_j^4/4 - 2\sum_j \sum_{k\neq j} \left([\mathbf{z}_k^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*)\right)^2 \\
&= 2r^2 - 2\sum_{j,k} \left([\mathbf{z}_k^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*)\right)^2 \\
&= 2r^2 - 2\sum_j \|[Z^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*)\|^2
\end{aligned}$$

By Lemma 7.3 we have for every $j = 1, \dots, K$

$$\|[Z^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*)\| \leq \|[Z^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*)\|_1 \leq 3.05\alpha^{-1/2} r_j^2,$$

therefore

$$\sum_j \|[Z^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*)\|^2 \leq 10\alpha^{-1}\sum_j r_j^4 \leq 10\alpha^{-1} r^4,$$

thus inequality follows. □

Lemma 7.5. *Let C, C' be such that $|C \triangle C'| = 1$. Then $\|\mathbf{z}_C - \mathbf{z}_{C'}\|^2 \leq \frac{2}{|C|\vee|C'|}$.*

Proof. Suppose, $|C'| > |C|$ then $C' = C \cup \{a\}$ and denoting $n = |C|$ we have

$$\|\mathbf{z}_C - \mathbf{z}_{C'}\|^2 = n\left(\sqrt{\frac{1}{n+1}} - \sqrt{\frac{1}{n}}\right)^2 + \frac{1}{n+1} = \frac{(\sqrt{n+1} - \sqrt{n})^2 + 1}{n+1} \leq \frac{2}{n+1}.$$

□

7.2 Proof of Theorem 3.6

The proof consists of several steps, each represented by a separate lemma.

Lemma 7.6. *Suppose, Assumption 1 holds and let $N \geq 2$. There is a constant $C = C(\gamma, L)$, so that if*

$$\frac{s\log N \log^2 T}{T p_{\min}^2} \leq \frac{1}{3},$$

then with probability at least $1 - 1/N$ and for with $\Delta_1 = C\sigma_{\max}\sqrt{\frac{\log N}{T p_{\min}^2}}$ the following inequalities take place for every $j = 1, \dots, K$

- $$\|\hat{A} - A\|_{\infty,\infty} \leq \Delta_1, \qquad \|\Sigma^{-1}_{\Lambda_j,\Lambda_j}(\hat{A}_{\Lambda_j,\cdot} - A_{\Lambda_j,\cdot})\|_{\infty,\infty} \leq \sigma_{\min}^{-1}\Delta_1; \tag{7.1}$$
- $$\|(\hat{A} - A)\mathbf{z}_j^*\|_\infty \leq \Delta_1, \qquad \|\Sigma^{-1}_{\Lambda_j,\Lambda_j}(\hat{A}_{\Lambda_j,\cdot} - A_{\Lambda_j,\cdot})\mathbf{z}_j^*\|_\infty \leq \sigma_{\min}^{-1}\Delta_1; \tag{7.2}$$
- $$\|\hat{\Sigma} - \Sigma\|_{\infty,\infty} \leq \Delta_1, \qquad \|(\hat{\Sigma}_{\Lambda_j,\cdot} - \Sigma_{\Lambda_j,\cdot})\mathbf{v}_j^*\|_\infty \leq \Delta_1; \tag{7.3}$$
- $$\|\Sigma^{-1}_{\Lambda_j,\Lambda_j}(\hat{\Sigma}_{\Lambda_j,\cdot} - \Sigma_{\Lambda_j,\cdot})\mathbf{v}_j^*\|_\infty \leq \sigma_{\min}^{-1}\Delta_1; \tag{7.4}$$

•

$$|||\hat{\Sigma}_{\Lambda_j,\Lambda_j} - \Sigma_{\Lambda_j,\Lambda_j}|||_{\mathsf{op}} \le \sqrt{s}\Delta_1. \tag{7.5}$$

Proof. By Theorem 3.5 for any pair $\mathbf{a}, \mathbf{b} \in \mathbb{R}^N$ with $\|\mathbf{a}\| \le 1$, $\|\mathbf{b}\| \le 1$ it holds with probability $\ge 1 - N^{-m}$,

$$|\mathbf{a}^\top(\hat{A} - A)\mathbf{b}| \le C\sigma_{\max}\left\{\sqrt{\frac{(m+1)\log N}{Tp_{\min}^2}} \bigvee \frac{(m+1)\log N \log T}{Tp_{\min}^2}\right\}.$$

Suppose for a moment that m is such that

$$\sqrt{\frac{(m+1)s\log N}{Tp_{\min}^2}}\log T \le 1, \tag{7.6}$$

so that we can neglect the second term. Set,

$$A_0 = \{(\mathbf{e}_i, \mathbf{e}_{i'}) :\ i, i' \le N\}, \qquad B_0 = \{(\mathbf{e}_i, \mathbf{z}_l^*) :\ i \le N, l \le K\},$$

as well as for every $j = 1, \dots, K$

$$A_j = \{(\sigma_{\min}\Sigma_{\Lambda_j,\Lambda_j}^{-1}\mathbf{e}_i, \mathbf{e}_{i'}) :\ i \in \Lambda_j, i' \le N\},$$
$$B_j = \{(\sigma_{\min}\Sigma_{\Lambda_j,\Lambda_j}^{-1}\mathbf{e}_i, \mathbf{z}_l^*) :\ i \in \Lambda_j, l \le K\}.$$

We have $|A_0| \le N^2, |B_0| \le NK$ and $|A_j| \le sN, |B_j| \le sK$ for $j = 1, \dots, N$, so since $s, K \le N$ together they have not more than $4N^3$ pairs of vectors $(\mathbf{a}, \mathbf{b})$, each having norm bounded by one. Taking a union bound, we have that the inequalities (7.1) and (7.2) hold with probability at least $1 - 4N^{3-m}$. By analogy, we can show that (7.3) and (7.4) hold with probability at least $1 - 4N^{3-m}$.

As for the last inequality, for every $j = 1, \dots, K$ pick $P_j = \sum_{i\in\Lambda_j} \mathbf{e}_i\mathbf{e}_i^\top$, i.e., projectors onto the subspace of vectors supported on Λ_j. Then by Theorem 3.4 it holds with probability at least $1 - KN^{-m}$ for every $j = 1, \dots, K$ (taking into account (7.6))

$$|||\hat{\Sigma}_{\Lambda_j,\Lambda_j} - \Sigma_{\Lambda_j,\Lambda_j}|||_{\mathsf{op}} = |||P_j(\hat{\Sigma} - \Sigma)P_j|||_{\mathsf{op}} \le C\sigma_{\max}\sqrt{\frac{s(m+1)\log N}{Tp_{\min}^2}}.$$

The total probability will be at least $1 - 8N^{3-m} - KN^{-m}$, which is at least $1 - 1/N$ whenever $m \ge 7$ and $N \ge 2$.

□

In the following, we apply the technique from Gribonval et al. (2015). Suppose that the LASSO solution $\hat{\mathbf{v}}_j$ for a given clustering $\mathcal{C}$ is not only supported exactly on Λ_j, but its signs are matching those of the true $\mathbf{v}_j^*$. Then, $\|\hat{\mathbf{v}}_j\|_1 = \bar{\mathbf{s}}_j^\top(\hat{\mathbf{v}}_j)_{\Lambda_j}$. Therefore, we can write

$$\begin{aligned}(\hat{\mathbf{v}}_j)_{\Lambda_j} &= \arg\min_{\mathbf{v}\in\mathbb{R}^{\Lambda_j}} \frac{1}{2}\mathbf{v}^\top\hat{\Sigma}_{\Lambda_j,\Lambda_j}\mathbf{v} - \mathbf{v}^\top\hat{A}_{\Lambda_j,\cdot}\mathbf{z}_j + \lambda\bar{\mathbf{s}}_j^\top\mathbf{v}\\ &= \hat{\Sigma}_{\Lambda_j,\Lambda_j}^{-1}(\hat{A}_{\Lambda_j,\cdot}\mathbf{z}_j - \lambda\bar{\mathbf{s}}_j),\end{aligned}$$

and plugging this solution into the risk function we get that $F_\lambda(\mathcal{C}) = \Phi_\lambda(\mathcal{C})$, where the

latter is defined explicitly

$$\Phi_\lambda(\mathcal{C}) = -\frac{1}{2}\sum_{j=1}^{K}(\hat{A}_{\Lambda_j,.}\mathbf{z}_j - \lambda\bar{\mathbf{s}}_j)^\top \hat{\Sigma}^{-1}_{\Lambda_j,\Lambda_j}(\hat{A}_{\Lambda_j,.}\mathbf{z}_j - \lambda\bar{\mathbf{s}}_j).$$

The next lemma shows that such representation takes place in the local vicinity of the true clustering $\mathcal{C}^*$.

Lemma 7.7. *Suppose, the inequalities* (7.1)–(7.5) *take place. Assume,*

$$s\Delta_1 \le 1/16, \qquad 12\Delta_1 \le \lambda \le \frac{\sigma_{\min}}{4}\tau_0 s^{-1}. \tag{7.7}$$

Then, for any $\mathcal{C} = (C_1, \dots, C_K)$ *satisfying*

$$\max_j \|\mathbf{z}_{C_j} - \mathbf{z}_{C_j^*}\| \le 0.3 \wedge 0.22\sqrt{\left(2\sigma_{\max}\alpha^{-1/2} + \sqrt{n^*}\Delta_1\right)^{-1}\lambda} \tag{7.8}$$

it holds

$$|||\hat{V}_{\lambda,\mathcal{C}} - V^*|||_{\mathsf{F}} \le 3\sigma_{\min}^{-1}\sqrt{Ks}\lambda,$$

and the equality $F_\lambda(\mathcal{C}) = \Phi_\lambda(\mathcal{C})$ *takes place.*

Proof. Taking into account $Z^\top Z = I_K$, it holds

$$\begin{aligned} R_\lambda(V;\mathcal{C}) =& \frac{1}{2}\mathrm{Tr}\left(V^\top\hat{\Sigma}V\right) - \mathrm{Tr}\left(V^\top\hat{A}Z\right) + \lambda\|V\|_{1,1} \\ =& \sum_{j=1}^{K}\frac{1}{2}\mathbf{v}_j^\top\hat{\Sigma}\mathbf{v}_j - \mathbf{v}_j^\top\hat{A}\mathbf{z}_j + \lambda\|\mathbf{v}_j\|_1, \end{aligned}$$

so that the optimization problem separates into K independent subproblems. Solving each of the problems

$$\frac{1}{2}\mathbf{v}_j^\top\hat{\Sigma}\mathbf{v}_j - \mathbf{v}_j^\top\hat{A}\mathbf{z}_j + \lambda\|\mathbf{v}_j\|_1 \to \min_{\mathbf{v}_j}$$

corresponds to Corollary B.3 with $\hat{D} = \hat{\Sigma}$ and $\hat{\mathbf{c}} = \hat{A}\mathbf{z}_j$, whereas the "true" version of the problem corresponds to $\bar{D} = \Sigma$ and $\bar{\mathbf{c}} = A\mathbf{z}_j^* = \Sigma(\Theta^*)^\top\mathbf{z}_j^* = \Sigma\mathbf{v}_j^*$. We need to control the differences between $\hat{\mathbf{c}}$ and $\bar{\mathbf{c}}$, and between $\hat{D}$ and $\bar{D}$. It holds,

$$\|\hat{A}\mathbf{z}_j - A\mathbf{z}_j^*\|_\infty \le \|A(\mathbf{z}_j - \mathbf{z}_j^*)\|_\infty + \|(\hat{A} - A)\mathbf{z}_j^*\|_\infty + \|(\hat{A} - A)(\mathbf{z}_j - \mathbf{z}_j^*)\|_\infty .$$

Since $A = \Sigma V^*[Z^*]^\top$, we bound the first term using Lemma 7.3

$$\|A(\mathbf{z}_j - \mathbf{z}_j^*)\|_\infty \le \|\Sigma V^*\|_{\infty,\infty}\|[Z^*]^\top(\mathbf{z}_j - \mathbf{z}_j^*)\|_1 \le 3.05\alpha^{-1/2}\|\Sigma V^*\|_{\infty,\infty}r_j^2.$$

The second term is bounded by Δ_1, whereas the fourth term satisfies

$$\|(\hat{A} - A)(\mathbf{z}_j - \mathbf{z}_j^*)\|_\infty \le \|\hat{A} - A\|_{\infty,\infty}\|\mathbf{z}_j - \mathbf{z}_j^*\|_1 \le 1.65\Delta_1\sqrt{n^*}r_j^2,$$

where we also used Lemma 7.2. Summing up,we get,

$$\|\hat{\mathbf{c}} - \mathbf{c}\|_\infty \le 1.65(2\sigma_{\max}\alpha^{-1/2} + \sqrt{n_j^*}\Delta_1)r_j^2 + \Delta_1 .$$

Similarly, we bound $\|\Sigma_{\Lambda_j,\Lambda_j}(\hat{\mathbf{c}}_{\Lambda_j} - \bar{\mathbf{c}}_{\Lambda_j})\|_\infty$ as follows

$$\begin{aligned}\|\Sigma^{-1}_{\Lambda_j,\Lambda_j}(\hat{A}_{\Lambda_j,\cdot}\mathbf{z}_j - A_{\Lambda_j,\cdot}\mathbf{z}^*_j)\|_\infty \leq& \|\Sigma^{-1}_{\Lambda_j,\Lambda_j}A(\mathbf{z}_j - \mathbf{z}^*_j)\|_\infty + \|\Sigma^{-1}_{\Lambda_j,\Lambda_j}(\hat{A}_{\Lambda_j,\cdot} - A_{\Lambda_j,\cdot})\mathbf{z}^*_j\|_\infty \\ &+ \|\Sigma^{-1}_{\Lambda_j,\Lambda_j}(\hat{A}_{\Lambda_j,\cdot} - A_{\Lambda_j,\cdot})(\mathbf{z}_j - \mathbf{z}^*_j)\|_\infty \\ \leq& \|\Sigma^{-1}_{\Lambda_j,\Lambda_j}A(\mathbf{z}_j - \mathbf{z}^*_j)\|_\infty + 1.65\sigma^{-1}_{\min}\Delta_1\sqrt{n^*}r_j^2 + \sigma^{-1}_{\min}\Delta_1 \\ \leq& 1.65\sigma^{-1}_{\min}(2\sigma_{\max}\alpha^{-1/2} + \sqrt{n^*_j}\Delta_1)r_j^2 + \sigma^{-1}_{\min}\Delta_1\end{aligned}$$

To sum up, Corollary B.3 is applied with

$$\begin{aligned}\delta_c =& 1.65(2\sigma_{\max}\alpha^{-1/2} + \sqrt{n^*}\Delta_1)r_j^2 + \Delta_1, \\ \delta'_c =& 1.65\sigma^{-1}_{\min}(2\sigma_{\max}\alpha^{-1/2} + \sqrt{n^*}\Delta_1)r_j^2 + \sigma^{-1}_{\min}\Delta_1 \\ \delta_D =& \Delta_1, \qquad \delta'_D = \Delta_1, \qquad \delta''_D = \sigma^{-1}_{\min}\Delta_1.\end{aligned}$$

It requires the conditions,

$$3\{1.65(2\sigma_{\max}\alpha^{-1/2} + \sqrt{n^*}\Delta_1)r_j^2 + 2\Delta_1\} \leq \lambda, \qquad s\Delta_1 \leq \frac{1}{16},$$

and due to the fact that $\|D^{-1}_{\Lambda_j,\Lambda_j}\|_{1,\infty} \leq \sqrt{s}|||D^{-1}_{\Lambda_j,\Lambda_j}|||_{\mathsf{op}}$ and Assumption 3.8,

$$2\sigma^{-1}_{\min}(1.65(2\sigma_{\max}\alpha^{-1/2} + \sqrt{n^*}\Delta_1)r_j^2 + 2\Delta_1 + \sqrt{s}\lambda) < \tau_0 s^{-1/2},$$

which are not hard to derive from the given inequalities. Together this yields that $\hat{\mathbf{v}}_j$ is supported on Λ_j and the solution satisfies

$$(\hat{\mathbf{v}}_j)_{\Lambda_j} = \hat{\Sigma}^{-1}_{\Lambda_j,\Lambda_j}\left(\hat{A}_{\Lambda_j,\cdot}\mathbf{z}_j - \lambda\mathbf{s}^*_j\right),$$

and the corresponding minimum is equal to

$$\frac{1}{2}\hat{\mathbf{v}}_j^\top\hat{\Sigma}\hat{\mathbf{v}}_j^\top - \hat{\mathbf{v}}_j^\top\hat{A}\mathbf{z}_j + \lambda(\hat{\mathbf{v}}_j)^\top_{\Lambda_j}\mathbf{s}^*_j = -\frac{1}{2}\left(\hat{A}_{\Lambda_j,\cdot}\mathbf{z}_j - \lambda\mathbf{s}^*_j\right)^\top\hat{\Sigma}^{-1}_{\Lambda_j,\Lambda_j}\left(\hat{A}_{\Lambda_j,\cdot}\mathbf{z}_j - \lambda\mathbf{s}^*_j\right).$$

Summing up, we get the corresponding expression for $F_\lambda(\mathcal{C})$. Moreover, we have

$$\begin{aligned}\|\hat{\mathbf{v}}_j - \mathbf{v}^*_j\| \leq& 2\sqrt{s}\left\{2\Delta_1 + 1.65(2\sigma_{\max}\alpha^{-1} + \sqrt{n^*}\Delta_1)r_j^2 + \lambda\right\} \\ \leq& 2\sigma^{-1}_{\min}\sqrt{s}\left(\frac{\lambda}{6} + \frac{1.65\lambda}{20} + \lambda\right) \\ \leq& 3\sigma^{-1}_{\min}\sqrt{s}\lambda,\end{aligned}$$

and together it provides a bound on $|||\hat{V}_{\lambda,\mathcal{C}} - V^*|||_{\mathsf{F}}$. □

Consider the function,

$$\bar{\Phi}_\lambda(\mathcal{C}) = -\frac{1}{2}\sum_{j=1}^k\left(A_{\Lambda_j,\cdot}\mathbf{z}_j - \lambda\mathbf{s}^*_j\right)^\top\Sigma^{-1}_{\Lambda_j,\Lambda_j}\left(A_{\Lambda_j,\cdot}\mathbf{z}_j - \lambda\mathbf{s}^*_j\right).$$

The following lemma shows how this function grows with $\mathcal{C}$ retreating from the true

clustering $\mathcal{C}^*$.

Lemma 7.8. *Suppose, $\mathcal{C}$ is a clustering such that $r = |||Z_{\mathcal{C}} - Z^*|||_{\mathsf{F}} \le 0.3$. Then,*

$$\bar{\Phi}_\lambda(\mathcal{C}) - \bar{\Phi}_\lambda(\mathcal{C}^*) \ge \frac{a_0}{2} r^2(1 - 10\alpha^{-1}r^2) - \lambda\sqrt{Ks}|||V^*|||_{\mathsf{F}} r.$$

Proof. Denoting $\bar{\Phi}_0(\mathcal{C}) = -\frac{1}{2}\sum_{j=1}^k \mathbf{z}_j^\top \hat{A}_{\Lambda_j,\cdot}^\top \hat{\Sigma}^{-1}_{\Lambda_j,\Lambda_j} \hat{A}_{\Lambda_j,\cdot}\mathbf{z}_j$ (which indeed corresponds to $\lambda = 0$), we have the decomposition

$$\bar{\Phi}_\lambda(\mathcal{C}) - \bar{\Phi}_\lambda(\mathcal{C}^*) = \bar{\Phi}_0(\mathcal{C}) - \bar{\Phi}_0(\mathcal{C}^*) - \lambda\sum_{j=1}^K [\mathbf{s}_j^*]^\top \Sigma^{-1}_{\Lambda_j,\Lambda_j} A_{\Lambda_j,\cdot}(\mathbf{z}_j - \mathbf{z}_j^*).$$

Let us first deal with the term $\bar{\Phi}_0(\mathcal{C}) - \bar{\Phi}_0(\mathcal{C}^*)$. Note that since $[\mathbf{v}_j^*]_{\Lambda_j} = \Sigma^{-1}_{\Lambda_j,\Lambda_j} A_{\Lambda_j,\cdot}\mathbf{z}_j^*$, we have

$$\bar{\Phi}_0(\mathcal{C}^*) = -\frac{1}{2}\sum_{j=1}^K [\mathbf{v}_j^*]^\top \Sigma \mathbf{v}_j^* = -\frac{1}{2}\mathrm{Tr}([V^*]^\top \Sigma V^*) = -\frac{1}{2}\mathrm{Tr}(\Theta^*\Sigma[\Theta^*]^\top).$$

whereas

$$\bar{\Phi}_0(\mathcal{C}) = \min_{V=[\mathbf{v}_1,\dots,\mathbf{v}_k]} \frac{1}{2}\mathrm{Tr}(V^\top \Sigma V) - \mathrm{Tr}(V^\top A Z_{\mathcal{C}})$$

where the minimum is taken s.t. the restrictions $\mathrm{supp}(\mathbf{v}_j) \subset \Lambda_j$. Dropping the restrictions we get,

$$\begin{aligned}
\bar{\Phi}_0(\mathcal{C}) - \bar{\Phi}_0(\mathcal{C}^*) &\ge \min_V \frac{1}{2}\mathrm{Tr}(V^\top \Sigma V) - \mathrm{Tr}(V^\top A Z_{\mathcal{C}}) + \frac{1}{2}\mathrm{Tr}(\Theta^*\Sigma[\Theta^*]^\top) \\
&= \min_V \frac{1}{2}|||Z_{\mathcal{C}} V^\top \Sigma^{1/2}|||_{\mathsf{F}}^2 - \mathrm{Tr}(Z_{\mathcal{C}} V^\top \Sigma[\Theta^*]^\top) + |||\Theta^*\Sigma^{1/2}|||_{\mathsf{F}}^2 \\
&= \min_V \frac{1}{2}|||(Z_{\mathcal{C}} V^\top - \Theta^*)\Sigma^{1/2}|||_{\mathsf{F}}^2.
\end{aligned}$$

It is not hard to calculate that the minimum is attained for $V = [\Theta^*]^\top Z_{\mathcal{C}}$ and therefore

$$\bar{\Phi}_0(\mathcal{C}) - \bar{\Phi}_0(\mathcal{C}^*) \ge \frac{1}{2}|||(Z_{\mathcal{C}} Z_{\mathcal{C}}^\top - I)\Theta^*\Sigma^{1/2}|||_{\mathsf{F}}^2 \ge \frac{a_0}{2}|||(Z_{\mathcal{C}} Z_{\mathcal{C}}^\top - I)Z^*|||_{\mathsf{F}}^2,$$

where the latter follows using $\Theta^* = Z^*[V^*]^\top$ and from the fact that $\lambda_{\min}([V^*]^\top \Sigma V^*) \ge \sigma_0$. Moreover,

$$\begin{aligned}
|||(Z_{\mathcal{C}} Z_{\mathcal{C}}^\top - I)Z^*|||_{\mathsf{F}}^2 &= \mathrm{Tr}((P_{\mathcal{C}} - I)P_{\mathcal{C}^*}(P_{\mathcal{C}} - I)) = \mathrm{Tr}(P_{\mathcal{C}^*}) - \mathrm{Tr}(P_{\mathcal{C}} P_{\mathcal{C}^*}) \\
&= \frac{1}{2}|||P_{\mathcal{C}} - P_{\mathcal{C}^*}|||_{\mathsf{F}}^2,
\end{aligned}$$

where we used the fact that $\mathrm{Tr}(P_{\mathcal{C}}) = \mathrm{Tr}(P_{\mathcal{C}^*}) = K$. It is left to recall the result of Lemma 7.4, so that we get

$$\bar{\Phi}_0(\mathcal{C}) - \bar{\Phi}_0(\mathcal{C}^*) \ge \frac{a_0 r^2}{2}(1 - 10\alpha^{-1}r^2).$$

As for the linear term, it holds

$$\left(\sum_{j=1}^{K}[\mathbf{s}_j^*]^\top \Sigma^{-1}_{\Lambda_j,\Lambda_j} A_{\Lambda_j,\cdot}(\mathbf{z}_j - \mathbf{z}_j^*)\right)^2 \le \left(\sum_{j=1}^{K}\|[\mathbf{s}_j^*]^\top \Sigma^{-1}_{\Lambda_j,\Lambda_j} A_{\Lambda_j,\cdot}\|^2\right) r^2$$

Since $A = \Sigma[\Theta^*]^\top$, we have $A^\top_{\Lambda_j,\cdot}\Sigma^{-1}_{\Lambda_j,\Lambda_j}\mathbf{s}_j^* = \Theta^*\Sigma_{\cdot,\Lambda_j}\Sigma^{-1}_{\Lambda_j,\Lambda_j}\mathbf{s}_j^*$. Denote, $\mathbf{x} = \Sigma_{\cdot,\Lambda_j}\Sigma^{-1}_{\Lambda_j,\Lambda_j}\mathbf{s}_j^*$, then we have $\mathbf{x}_{\Lambda_j} = \mathbf{s}_j$ and $\|\mathbf{x}_{\Lambda_j}\|_\infty = 1$. Moreover, by the ERC property

$$\|\mathbf{x}_{\Lambda_j^c}\|_\infty = \|\Sigma_{\Lambda_j^c,\Lambda_j}\Sigma^{-1}_{\Lambda_j,\Lambda_j}\mathbf{s}_j\|_\infty \le \|\Sigma_{\Lambda_j^c,\Lambda_j}\Sigma^{-1}_{\Lambda_j,\Lambda_j}\|_{1,\infty} \le 1/2.$$

We have

$$\|A^\top_{\Lambda_j,\cdot}\Sigma^{-1}_{\Lambda_j,\Lambda_j}\mathbf{s}_j^*\|^2 = \|\sum \mathbf{z}_j^*[\mathbf{v}_j^*]^\top\mathbf{x}\|^2 = \sum_{k=1}^{K}|[\mathbf{v}_k^*]^\top\mathbf{x}|^2,$$

where, since $\mathbf{v}_k^*$ is supported on Λ_k of size at most s,

$$|[\mathbf{v}_k^*]^\top\mathbf{x}| \le \|\mathbf{v}_k^*\|_1\|\mathbf{x}\|_\infty \le \sqrt{s}\|\mathbf{v}_k^*\|.$$

Summing up, we get $\|A^\top_{\Lambda_j,\cdot}\Sigma^{-1}_{\Lambda_j,\Lambda_j}\mathbf{s}_j^*\|^2 \le s|||V^*|||_{\mathsf{F}}^2$, so that

$$\left|\sum_{j=1}^{K}[\mathbf{s}_j^*]^\top \Sigma^{-1}_{\Lambda_j,\Lambda_j} A_{\Lambda_j,\cdot}(\mathbf{z}_j - \mathbf{z}_j^*)\right| \le \sqrt{Ks}|||V^*|||_{\mathsf{F}} r.$$

The lemma now follows from the two terms put together. □

The next step is to bound the difference $\Phi_\lambda(\mathcal{C}) - \bar{\Phi}_\lambda(\mathcal{C})$ uniformly in the neighbourhood of $\mathcal{C}^*$.

Lemma 7.9. *Suppose that the inequalities* (7.1)–(7.5) *hold and let*

$$\Delta_1 \le \sigma_{\min}/(2\sqrt{s}) \vee \frac{\lambda}{12}, \qquad \sigma_{\max}/\sigma_{\min} \le n^*, \qquad \lambda \le \sigma_{\min}s^{-1}$$

Let some $r \le 0.3$ *satisfies* $\sqrt{sn^*}\Delta_1 r^2 \le \sigma_{\max}$. *Then,*

$$\sup_{|||Z-Z^*|||_{\mathsf{F}}\le r} |\Phi_\lambda(\mathcal{C}) - \bar{\Phi}_\lambda(\mathcal{C}) - \Phi_\lambda(\mathcal{C}^*) + \bar{\Phi}_\lambda(\mathcal{C}^*)|$$
$$\le 4\left(\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2\sqrt{s}|||V^*|||_{\mathsf{F}} + \frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{K}\right)\Delta_1 r + 16\frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{sn^*}\Delta_1 r^2.$$

Proof. Denote,

$$\tilde{\Phi}_\lambda(\mathcal{C}) = -\frac{1}{2}\sum_{j=1}^{K}\left(A_{\Lambda_j,\cdot}\mathbf{z}_j - \lambda\mathbf{s}_j^*\right)^\top\hat{\Sigma}^{-1}_{\Lambda_j,\Lambda_j}\left(A_{\Lambda_j,\cdot}\mathbf{z}_j - \lambda\mathbf{s}_j^*\right),$$

so that we have

$$
\begin{aligned}
&|\tilde{\Phi}_\lambda(\mathcal{C}) - \bar{\Phi}_\lambda(\mathcal{C}) - \tilde{\Phi}_\lambda(\mathcal{C}^*) + \bar{\Phi}_\lambda(\mathcal{C}^*)| \\
&\qquad \le \frac{1}{2}\sum_{j=1}^{K} \left| \left(A_{\Lambda_j,\cdot}(\mathbf{z}_j + \mathbf{z}_j^*) - 2\lambda \mathbf{s}_j^*\right)^\top (\hat{\Sigma}^{-1}_{\Lambda_j,\Lambda_j} - \Sigma^{-1}_{\Lambda_j,\Lambda_j}) A_{\Lambda_j,\cdot}(\mathbf{z}_j - \mathbf{z}_j^*) \right|
\end{aligned}
$$

First of all, due to (7.5) it holds,

$$
|||\hat{\Sigma}^{-1}_{\Lambda_j,\Lambda_j} - \Sigma^{-1}_{\Lambda_j,\Lambda_j}|||_{\mathsf{op}} \le \frac{\sigma_{\min}^{-2}\sqrt{s}\Delta_1}{1 - \sigma_{\min}^{-1}\sqrt{s}\Delta_1} \le 2\sigma_{\min}^{-2}\sqrt{s}\Delta_1.
$$

Since $A = \Sigma[\Theta^*]^\top$, we have

$$
\begin{aligned}
\|A_{\Lambda_j,\cdot}(\mathbf{z}_j - \mathbf{z}_j^*)\| &\le \sigma_{\max} r_j \\
\|A_{\Lambda_j,\cdot}(\mathbf{z}_j + \mathbf{z}_j^*) - 2\lambda \mathbf{s}_j^*\| &\le \sigma_{\max}(2\|\mathbf{v}_j^*\| + r_j) + 2\lambda\sqrt{s}.
\end{aligned}
$$

Then by Cauchy-Schwartz,

$$
\begin{aligned}
|\tilde{\Phi}_\lambda(\mathcal{C}) - \bar{\Phi}_\lambda(\mathcal{C}) - \tilde{\Phi}_\lambda(\mathcal{C}^*) + \bar{\Phi}_\lambda(\mathcal{C}^*)| \le& \sigma_{\min}^{-2}\sqrt{s}\Delta_1 \left(\sum_{j=1}^{K} \sigma_{\max} r_j \left\{ \sigma_{\max}(2\|\mathbf{v}_j\| + r_j) + 2\lambda\sqrt{s} \right\} \right) \\
\le& 2\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2 \sqrt{s}|||V^*|||_{\mathsf{F}}\Delta_1 r + 2\frac{\sigma_{\max}}{\sigma_{\min}^2}\lambda s\sqrt{K}\Delta_1 r \\
&+ \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2 \sqrt{s}\Delta_1 r^2.
\end{aligned}
$$

Going further,

$$
\Phi_\lambda(\mathcal{C}) - \tilde{\Phi}_\lambda(\mathcal{C}) = -\frac{1}{2}\sum_{j=1}^{K} \left((A_{\Lambda_j,\cdot} + \hat{A}_{\Lambda_j,\cdot})\mathbf{z}_j - 2\lambda \mathbf{s}_j^* \right)^\top \hat{\Sigma}^{-1}_{\Lambda_j,\Lambda_j}(\hat{A}_{\Lambda_j,\cdot} - A_{\Lambda_j,\cdot})\mathbf{z}_j,
$$

which implies that

$$
\begin{aligned}
&|\Phi_\lambda(\mathcal{C}) - \tilde{\Phi}_\lambda(\mathcal{C}) - \Phi_\lambda(\mathcal{C}^*) + \tilde{\Phi}_\lambda(\mathcal{C}^*)| \\
&\qquad \le \frac{1}{2}\sum_{j=1}^{K} \left| \left((A_{\Lambda_j,\cdot} + \hat{A}_{\Lambda_j,\cdot})(\mathbf{z}_j - \mathbf{z}_j^*) \right)^\top \hat{\Sigma}^{-1}_{\Lambda_j,\Lambda_j}(\hat{A}_{\Lambda_j,\cdot} - A_{\Lambda_j,\cdot})\mathbf{z}_j \right| \\
&\qquad \frac{1}{2}\sum_{j=1}^{K} \left| \left((A_{\Lambda_j,\cdot} + \hat{A}_{\Lambda_j,\cdot})\mathbf{z}_j^* - 2\lambda \mathbf{s}_j^* \right)^\top \hat{\Sigma}^{-1}_{\Lambda_j,\Lambda_j}(\hat{A}_{\Lambda_j,\cdot} - A_{\Lambda_j,\cdot})(\mathbf{z}_j - \mathbf{z}_j^*) \right|
\end{aligned} \tag{7.9}
$$

First notice, that due to Lemma 7.2 and (7.1) it holds,

$$
\begin{aligned}
\|(\hat{A}_{\Lambda_j,\cdot} - A_{\Lambda_j,\cdot})(\mathbf{z}_j - \mathbf{z}_j^*)\| &\le \sqrt{s}\|\hat{A}_{\Lambda_j,\cdot} - A_{\Lambda_j,\cdot}\|_{\infty,\infty}\|\mathbf{z}_j - \mathbf{z}_j^*\|_1 \\
&\le 1.65\sqrt{sn^*}\Delta_1 r_j^2.
\end{aligned}
$$

Therefore, it follows

$$\|(\hat{A}_{\Lambda_j,\cdot} + A_{\Lambda_j,\cdot})(\mathbf{z}_j - \mathbf{z}_j^*)\| \le 2\sigma_{\max} r_j + 1.65\sqrt{sn^*}\Delta_1 r_j^2.$$

Moreover, using (7.2) we get

$$\|(\hat{A}_{\Lambda_j,\cdot} - A_{\Lambda_j,\cdot})\mathbf{z}_j\| \le \Delta_1 + 1.65\sqrt{sn^*}\Delta_1 r_j^2$$
$$\|(\hat{A}_{\Lambda_j,\cdot} + A_{\Lambda_j,\cdot})\mathbf{z}_j^* - 2\lambda\mathbf{s}_j^*\| \le 2\sigma_{\max}\|\mathbf{v}_j\| + \Delta_1 + 2\lambda\sqrt{s}.$$

and we also have $\|\!|\hat{\Sigma}^{-1}_{\Lambda_j,\Lambda_j}|\!\|_{\mathsf{op}} \le 2\sigma_{\min}^{-1}$ due to the condition $\sigma_{\min}^{-1}\sqrt{s}\Delta_1 \le 1/2$. Thus we get that the first sum of (7.9) is bounded by

$$\sigma_{\min}^{-1}\sum_{j=1}^{K}\left(2\sigma_{\max}r_j + 1.65\sqrt{sn^*}\Delta_1 r_j^2\right)\left(\Delta_1 + 1.65\sqrt{sn^*}\Delta_1 r_j^2\right)$$
$$\le 2\frac{\sigma_{\max}}{\sigma_{\min}}\Delta_1\sqrt{K}r + 1.65\sigma_{\min}^{-1}\sqrt{sn^*}\Delta_1^2 r^2 + 3.3\frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{sn^*}\Delta_1 r^3 + 2.8\sigma_{\min}^{-1}sn^*\Delta_1^2 r^4,$$

while the second sum is bounded by

$$\sigma_{\min}^{-1}\sum_{j=1}^{K}\left(2\sigma_{\max}\|\mathbf{v}_j^*\| + \Delta_1 + 2\lambda\sqrt{s}\right)\left(1.65\sqrt{sn^*}\Delta_1 r_j^2\right)$$
$$\le \frac{1.65}{\sigma_{\min}}\left(\sigma_{\max}\sqrt{sn^*} + \sqrt{sn^*}\Delta_1 + 2\lambda s\sqrt{n^*}\right)\Delta_1 r^2$$
$$\le \frac{3.3}{\sigma_{\min}}\left(\sigma_{\max}\sqrt{sn^*} + \lambda s\sqrt{n^*}\right)\Delta_1 r^2$$

where we used the fact that $\max_j\|\mathbf{v}_j^*\| \le \|\!|V^*|\!\|_{\mathsf{op}} = \|\!|\Theta^*|\!\|_{\mathsf{op}} < 1$ together with the condition of the lemma $\Delta_1 \le \sigma_{\max}$. Combining all the bounds we get

$$\begin{aligned}
&|\Phi_\lambda(\mathcal{C}) - \bar{\Phi}_\lambda(\mathcal{C}) - \Phi_\lambda(\mathcal{C}^*) + \bar{\Phi}_\lambda(\mathcal{C}^*)| \\
&\quad\le 2\left\{\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2\sqrt{s}\|\!|V^*|\!\|_{\mathsf{F}} + 2\frac{\sigma_{\max}}{\sigma_{\min}^2}\lambda s\sqrt{K} + 2\frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{K}\right\}\Delta_1 r \\
&\qquad + \left\{3.3\frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{sn^*} + 3.3\sigma_{\min}^{-1}\lambda s\sqrt{n^*} + 1.65\sigma_{\min}^{-1}\sqrt{sn^*}\Delta_1 + \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2\sqrt{s}\right\}\Delta_1 r^2 \\
&\qquad + 3.3\frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{sn^*}\Delta_1 r^3 \\
&\qquad + 2.8\sigma_{\min}^{-1}sn^*\Delta_1^2 r^4,
\end{aligned}$$

where by $r \le 0.3$ and $\sqrt{sn^*}\Delta_1 \le \sigma_{\max}$ we can neglect the third and the fourth power, respectively, and thus the required bound follows. □

Lemma 7.10. *There are numerical constants $c, C > 0$ such that the following holds. Suppose, the inequalities take place:*

$$\sqrt{\frac{sn^*\log N}{Tp_{\min}^2}} \le c\frac{a_0\sigma_{\min}}{\sigma_{\max}^2}, \qquad n^* \ge \sigma_{\max}/\sigma_{\min}. \tag{7.10}$$

Let $C\sigma_{\max}\sqrt{\frac{\log N}{Tp_{\min}^2}} \le \lambda \le c\sigma_{\min}\tau_0 s^{-1}$*, and set*

$$\bar{r} = 0.3 \wedge 0.18\sqrt{\alpha} \wedge 0.22\sqrt{\left(2\sigma_{\max}\alpha^{-1/2} + \sqrt{n^*}\Delta_1\right)^{-1}\lambda}.$$

Then under the inequalities (7.1)–(7.5) *the clustering*

$$\hat{\mathcal{C}} = \arg\min_{|||Z_{\mathcal{C}} - Z^*|||_{\mathsf{F}} \le r_{\max}} F_\lambda(\mathcal{C})$$

satisfies

$$|||Z_{\hat{\mathcal{C}}} - Z^*|||_{\mathsf{F}} \le \frac{C}{a_0}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2 \lambda K\sqrt{s}\,.$$

Proof. It is not hard to see that for $\Delta_1 = \sqrt{\frac{\log N}{Tp_{\min}^2}}$ the inequalities required by Lemmas 7.7–7.9 are satisfied for $r \le \bar{r}$ due to (7.10) and conditions on λ and $\bar{r}$. Since obviously $\hat{\mathcal{C}}$ satisfies $F_\lambda(\hat{\mathcal{C}}) \le F_\lambda(\mathcal{C}^*)$, we have for $r = |||Z_{\hat{\mathcal{C}}} - Z_{\mathcal{C}^*}|||_{\mathsf{F}} \le r_{\max}$

$$\begin{aligned}
F_\lambda(\hat{\mathcal{C}}) - F_\lambda(\mathcal{C}^*) \ge& \bar{\Phi}_\lambda(\mathcal{C}) - \bar{\Phi}_\lambda(\mathcal{C}) - |F_\lambda(\mathcal{C}) - \bar{\Phi}_\lambda(\mathcal{C}) - F_\lambda(\mathcal{C}^*) + \bar{\Phi}_\lambda(\mathcal{C}^*)| \\
\ge& \frac{a_0 r^2}{2}\left(1 - 10\alpha^{-1}r^2\right) - \lambda\sqrt{Ks}|||V^*|||_{\mathsf{F}} r \\
&- 4\left\{\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2\sqrt{s}|||V^*|||_{\mathsf{F}} + \frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{K}\right\}\Delta_1 r - 15\frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{sn^*}\Delta_1 r^2 \\
=& \frac{a_0 r^2}{2}\left(1 - 10\alpha^{-1}r^2 - \frac{30}{a_0}\frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{sn^*}\Delta_1\right) \\
&- \lambda\sqrt{Ks}|||V^*|||_{\mathsf{F}} r - 4\left\{\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2\sqrt{s}|||V^*|||_{\mathsf{F}} + \frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{K}\right\}\Delta_1 r\,.
\end{aligned}$$

Since $\bar{r} \le 0.2\sqrt{\alpha}$ implies $10\alpha^{-1}r^2 \le \frac{1}{3}$, it holds by (7.10)

$$1 - 10\alpha^{-1}r^2 - \frac{30}{a_0}\frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{sn^*}\Delta_1 \ge \frac{1}{2}.$$

Therefore, after dividing by r, we get that such optimal clustering must satisfy

$$\frac{a_0}{4}r \le \lambda\sqrt{Ks}|||V^*|||_{\mathsf{F}} + 4\left\{\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2\sqrt{s}|||V^*|||_{\mathsf{F}} + \frac{\sigma_{\max}}{\sigma_{\min}}\sqrt{K}\right\}\Delta_1.$$

Recalling that $|||V^*|||_{\mathsf{F}} \le \sqrt{K}$, $\Delta_1 = C\sigma_{\max}\sqrt{\frac{\log N}{Tp_{\min}^2}}$, and $\Delta_2 = C\sqrt{\frac{s\log N}{Tp_{\min}^2}}$ yields the result. □

Now we are ready to finalize the proof of Theorem 3.6. Firstly, we need to show that the clustering $\hat{\mathcal{C}}$ from the lemma above is locally optimal. By Lemma 7.5, any neighbouring to it clustering $\mathcal{C}'$ satisfies $|||Z_{\mathcal{C}'} - Z_{\hat{\mathcal{C}}}|||_{\mathsf{F}} \le \frac{2}{\sqrt{\alpha N/K}}$. Therefore,

$$|||Z_{\mathcal{C}'} - Z_{\mathcal{C}^*}|||_{\mathsf{F}} \le \frac{C}{a_0}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2 \lambda K\sqrt{s} + 2\alpha^{-1/2}\sqrt{\frac{K}{N}}\,,$$

and it is enough to check that this value is at most $\bar{r}$. We check that each of the terms is at most $\bar{r}/2$. For the first one, it is sufficient to have

$$\frac{C}{a_0}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2 \alpha^{-1/2}\lambda K\sqrt{s} \le 0.09,$$
$$\frac{C^2}{a_0^2}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^4 \lambda\left(2\sigma_{\max}\alpha^{-1/2}+\sqrt{n^*}\Delta_1\right)K^2 s \le 0.012,$$

and both are satisfied due to the upper bound $\lambda \le c\kappa^{-4}(a_0^2/\sigma_{\max})K^{-2}s^{-1}$ and the requirement $\sqrt{\frac{sn^*\log N}{Tp_{\min}^2}} \le c$. For the second term we need

$$\alpha^{-1}\frac{K}{N} \le 0.008\alpha, \qquad \alpha^{-1}\left(2\sigma_{\max}\alpha^{-1/2}+\sqrt{n^*}\Delta_1\right)\frac{K}{N} \le \lambda,$$

both are satisfied once $N \ge C\alpha^2 K$ and $\lambda \ge C\sigma_{\max}\alpha^{-3/2}\frac{K}{N}$.

Moreover, by Lemma 7.7 we have for $\hat{\Theta} = Z_{\hat{\mathcal{C}}}\hat{V}_{\hat{\mathcal{C}},\lambda}$

$$\begin{aligned}\|\!|\hat{\Theta}-\Theta^*|\!\|_{\mathsf{F}} &\le \|\!|Z_{\hat{\mathcal{C}}}(\hat{V}_{\hat{\mathcal{C}},\lambda}-V^*)^\top|\!\|_{\mathsf{F}} + \|\!|(Z_{\hat{\mathcal{C}}}-Z^*)V^*|\!\|_{\mathsf{F}} \\ &\le 3\sigma_{\min}^{-1}\sqrt{Ks}\lambda + \frac{C}{a_0}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^2 \gamma K\sqrt{s}\lambda,\end{aligned}$$

which finishes the proof.

References

Avery, C. N., Chevalier, J. A., and Zeckhauser, R. J. (2016). The "CAPS" Prediction System and Stock Market Returns. *Review of Finance*, 20(4):1363–1381.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Bickel, P. J., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Brody, S. and Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: Using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 562–570.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 10–17.

Chen, C. Y.-H., Després, R., Guo, L., and Renault, T. (2019a). What makes cryptocurrencies special? Investor sentiment and price predictability during the bubble. *IRTG 1792 Discussion Paper 2019-016, submitted to Management Science*.

Chen, C. Y.-H., Härdle, W. K., and Okhrin, Y. (2019b). Tail event driven networks of SIFIs. *Journal of Econometrics*, 208(1):282–298.

Chen, S. and Schienle, M. (2019). Pre-screening and reduced rank regression for high-dimensional cointegration. *KIT working paper*.

Chen, Y., Trimborn, S., and Zhang, J. (2018). Discover Regional and Size Effects in Global Bitcoin Blockchain via Sparse-Group Network AutoRegressive Modeling. *Available at SSRN: https://ssrn.com/abstract=3245031*.

Chernozhukov, V., Härdle, W. K., Huang, C., and Wang, W. (2018). LASSO-Driven Inference in Time and Space. *Annals of Statistics, revise & resubmit*.

Čížek, P., Härdle, W., and Spokoiny, V. (2009). Adaptive pointwise estimation in time-inhomogeneous conditional heteroscedasticity models. *The Econometrics Journal*, 12(2):248–271.

Deng, S., Sinha, A. P., and Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems*, 94:65–76.

Diebold, F. X. and Yılmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 182(1):119–134.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521.

Gribonval, R., Jenatton, R., and Bach, F. (2015). Sparse and spurious: dictionary learning with noise and outliers. *IEEE Transactions on Information Theory*, 61(11):6298–6319.

Gudmundsson, G. and Brownlees, C. T. (2018). Community Detection in Large Vector Autoregressions. *Available at SSRN: https://ssrn.com/abstract=3072985*.

Han, F., Lu, H., and Liu, H. (2015). A Direct Estimation of High Dimensional Stationary Vector Autoregressions. *The Journal of Machine Learning Research*, 16(1):3115–3150.

Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):6 pp.

Kim, S.-H. and Kim, D. (2014). Investor sentiment from internet message postings and the predictability of stock returns. *Journal of Economic Behavior & Organization*, 107, Part B:708–729.

Klochkov, Y. and Zhivotovskiy, N. (2018). Uniform Hanson-Wright type concentration inequalities for unbounded entries via the entropy method. *arXiv preprint arXiv:1812.03548*.

Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133.

Le Gouic, T. and Paris, Q. (2018). A notion of stability for k-means clustering. *Electronic Journal of Statistics*, 12(2):4239–4263.

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Loughran, T. and McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54(4):1187–1230.

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058.

Melnyk, I. and Banerjee, A. (2016). Estimating structured vector autoregressive models. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 830–839.

Mihoci, A., Althof, M., Härdle, W. K., and Chen, C. Y.-H. (2019). FRM Financial Risk Meter. *Empirical Economics, forthcoming*.

Moon, H. R. and Weidner, M. (2018). Nuclear norm regularized estimation of panel regression models. *arXiv preprint arXiv:1810.10987*.

Rakhlin, A. and Caponnetto, A. (2007). Stability of k-means clustering. In *Proceedings of the 21th Annual Conference on Neural Information Processing Systems*, pages 1121–1128.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. *Journal of Banking & Finance*, 84:25–40.

Rohe, K., Qin, T., and Yu, B. (2016). Co-clustering directed graphs to discover asymmetries and directional communities. In *Proceedings of the National Academy of Sciences*, volume 113, pages 12679–12684.

Shindler, M., Wong, A., and Meyerson, A. W. (2011). Fast and Accurate k-means For Large Datasets. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, pages 2375–2383.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., and Welpe, I. M. (2014). Tweets and Trades: the Information Content of Stock Microblogs. *European Financial Management*, 20(5):926–957.

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051.

Udell, M., Horn, C., Zadeh, R., and Boyd, S. (2016). Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):57–83.

Zhu, X. and Pan, R. (2018). Grouped Network Vector Autoregression. *Statistica Sinca, to appear*.

Zhu, X., Pan, R., Li, G., Liu, Y., and Wang, H. (2017). Network vector autoregression. *The Annals of Statistics*, 45(3):1096–1123.

Zhu, X., Wang, W., Wang, H., and Härdle, W. K. (2019). Network quantile autoregression. *Journal of Econometrics*, 212:345–358.

A Proof of Theorems 3.4 and 3.5

Recall that we have a time series,

$$Y_t = \sum_{k\geq 0} \Theta^k W_{t-k}, \qquad t \in \mathbb{Z}, \tag{A.1}$$

where $W_t \in \mathbb{R}^N$, $t \in \mathbb{Z}$ are independent vectors with $\mathsf{E}W_t = 0$ and $\mathrm{Var}(W_t) = S$. We also have $|||\Theta|||_{\mathsf{op}} \leq \gamma$ for some $\gamma < 1$, and the covariance $\Sigma = \mathrm{Var}(Y_t)$ reads as

$$\Sigma = \sum_{k\geq 0} \Theta^k S [\Theta^k]^\top.$$

We have the observations

$$Z_t = (\delta_{1t} Y_{1t}, \ldots, \delta_{Nt} Y_{Nt})^\top, \qquad t = 1, \ldots, T, \tag{A.2}$$

where $\delta_{it} \sim \mathrm{Be}(p_i)$ are independent Bernoulli random variables for every $i = 1, \ldots, N$ and $t = 1, \ldots, T$ and some $p_i \in (0, 1]$.

The proofs of both statements are based on the following version of the Bernstein matrix inequality, which does not require bounded summands. Recall, that for a random variable $X \in \mathbb{R}$ the value

$$\|X\|_{\psi_j} = \inf\left\{ C > 0 : \mathsf{E}\exp\left(\left|\frac{X}{C}\right|^j\right) \leq 2 \right\}$$

denotes a ψ_j-norm. For $j = 1$ the norm is referred to as *subexponential* and for $j = 2$ as *subgaussian*.

Theorem A.1 (Klochkov and Zhivotovskiy (2018), Proposition 4.1)**.** *Suppose, the matrices A_t for $t = 1, \ldots, T$ are independent and let $M = \max_t \big\| |||A_t|||_{\mathsf{op}} \big\|_{\psi_1}$ is finite. Then, $S_T = \sum_{t=1}^T A_t$ satisfies for any $u \geq 1$*

$$\mathsf{P}\left[|||S_T - \mathsf{E}S_T|||_{\mathsf{op}} > C\left\{ \sqrt{\sigma^2(\log N + u)} + M \log T (\log N + u) \right\} \right] \leq e^{-u},$$

where $\sigma^2 = ||| \sum_{t=1}^T \mathsf{E}A_t^\top A_t |||_{\mathsf{op}} \vee ||| \sum_{t=1}^T \mathsf{E}A_t A_t^\top |||_{\mathsf{op}}$ and C is an absolute constant.

Let $\boldsymbol{\delta}_t = (\delta_{t1}, \dots, \delta_{tN})^\top$ denotes the vector with Bernoulli variables from above corresponding to the time point t. In what follows we consider the following matrices,

$$A^{k,j}_{t,t'} = \mathrm{diag}\{\boldsymbol{\delta}_t\}\Theta^k W_{t-k} W^\top_{t'-j}[\Theta^j]^\top \mathrm{diag}\{\boldsymbol{\delta}_{t'}\},$$

so that since $Z_t = \sum_{k\geq 0} \mathrm{diag}\{\boldsymbol{\delta}_t\}\Theta^k W_{t-k}$, we have

$$Z_t Z_t^\top = \sum_{k,j\geq 0} \mathrm{diag}\{\boldsymbol{\delta}_t\}\Theta^k W_{t-k} W^\top_{t-j}[\Theta^j]^\top \mathrm{diag}\{\boldsymbol{\delta}_t\} = \sum_{k,j\geq 0} A^{k,j}_{t,t}.$$

Therefore, the decomposition takes place

$$\Sigma^* = \sum_{k,j\geq 0} S_{k,j}, \qquad S_{k,j} = \frac{1}{T}\sum_{t=1}^{T} A^{k,j}_{t,t}, \tag{A.3}$$

and we shall analyze the sum $S_{k,j}$ for every pair of $k, j \geq 0$ separately. We first introduce two technical lemmas. In what follows we assume w.l.o.g. that $|||S|||_{\mathsf{op}} = 1$, since if we scale it, all the covariances and estimators scale correspondingly.

Lemma A.2. *Under the assumptions of Theorem 3.4 it holds,*

$$\begin{aligned}
\|\,|||P\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Diag}(A^{k,j}_{t,t'})Q|||_{\mathsf{op}}\|_{\psi_1} &\leq C p_{\min}^{-1}\sqrt{M_1 M_2}\gamma^{k+j},\\
\|\,|||P\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Off}(A^{k,j}_{t,t'})\mathrm{diag}\{\mathbf{p}\}^{-1}Q|||_{\mathsf{op}}\|_{\psi_1} &\leq C p_{\min}^{-2}\sqrt{M_1 M_2}\gamma^{k+j},
\end{aligned}$$

with some $C = C(L) > 0$.

Proof. Denote for simplicity $\mathbf{x} = \Theta^k W_{t-k}$, $\mathbf{y} = \Theta^j W_{t'-j}$, as well as $\mathbf{x}^\delta = \mathrm{diag}\{\boldsymbol{\delta}_t\}\mathbf{x}$, $\mathbf{y}^\delta = \mathrm{diag}\{\boldsymbol{\delta}_t\}\mathbf{y}$, such that $A^{k,j}_{t,t'} = \mathbf{x}^\delta[\mathbf{y}^\delta]^\top$. Since W_t are subgaussian and $|||\Theta^k S \Theta^k|||_{\mathsf{op}} \leq \gamma^{2k}$, we have for any $\mathbf{u} \in \mathbb{R}^N$

$$\log \mathsf{E}\exp(\mathbf{u}^\top \mathbf{x}) \leq C'\gamma^{2k}\|\mathbf{u}\|^2, \tag{A.4}$$

and since δ_t takes values in $[0,1]^N$, same takes place for $\mathbf{x}^{\boldsymbol{\delta}}$. By Theorem 2.1 in Hsu et al. (2012) it holds for any matrix A and vector $\mathbf{u} \in \mathbb{R}^N$,

$$\|\,|||A\mathbf{x}^\delta|||\,\|_{\psi_2} \leq C''\gamma^k |||A|||_{\mathsf{F}}, \qquad \|\mathbf{u}^\top \mathbf{x}^\delta\|_{\psi_2} \leq C''\gamma^k\|\mathbf{u}\|, \tag{A.5}$$

and, similarly,

$$\|\,|||A\mathbf{y}^\delta|||\,\|_{\psi_2} \leq C''\gamma^j |||A|||_{\mathsf{F}}, \qquad \|\mathbf{u}^\top \mathbf{y}^\delta\|_{\psi_2} \leq C''\gamma^j\|\mathbf{u}\|.$$

We first deal with the diagonal term. Let $P = \sum_{i=1}^{M_1}\mathbf{u}_j\mathbf{u}_j^\top$ be its eigen-decomposition with $\|\mathbf{u}_j\| = 1$, then

$$\begin{aligned}
\|\,|||P\mathrm{diag}(\mathbf{x}^\delta)|||_{\mathsf{op}}\|^2_{\psi_2} =& \|\,|||\mathrm{diag}(\mathbf{x}^\delta)P\mathrm{diag}(\mathbf{x}^\delta)|||_{\mathsf{op}}\|_{\psi_1} \leq \sum_{j=1}^{M_1}\|\,|||\mathrm{diag}(\mathbf{x}_\delta)\mathbf{u}_j\mathbf{u}_j^\top\mathrm{diag}(\mathbf{x}^\delta)|||_{\mathsf{op}}\|_{\psi_1}\\
=& \sum_{j=1}^{M_1}\|\,|||\mathrm{diag}(\mathbf{u}_j)\mathbf{x}^\delta|||\,\|^2_{\psi_2},
\end{aligned}$$

where each term in the latter is bounded by γ^{2k} due the fact that $|||\mathrm{diag}(\mathbf{u}_j)|||_{\mathsf{F}} = 1$.

Summing up and taking square root, we arrive at $\big\|\|\|P\mathrm{diag}(\mathbf{x}^\delta)\|\|\|_{\mathsf{op}}\big\|_{\psi_2} \le \sqrt{C''M_1}\gamma^k$. Taking into account similar bound for $Q\mathrm{diag}(\mathbf{y}^\delta)$, we have by Hölder inequality

$$\begin{aligned}\|\|\|P\mathrm{diag}\{\delta\}^{-1}\mathrm{diag}(\mathbf{x}^\delta)\mathrm{diag}(\mathbf{y}^\delta)Q\|\|\|_{\mathsf{op}}\|_{\psi_1} \le& p_{\min}^{-1}\|\|\|P\mathrm{diag}(\mathbf{x}^\delta)\|\|\|_{\mathsf{op}}\big\|_{\psi_2}\|\|\|Q\mathrm{diag}(\mathbf{y}^\delta)\|\|\|_{\mathsf{op}}\|_{\psi_2}\\ \le& C''\sqrt{M_1M_2}\gamma^{k+j},\end{aligned}$$

which yields the bound for the diagonal. As for the off-diagonal, consider first the whole matrix,

$$\|\|\|P\mathbf{x}^\delta[\mathbf{y}^\delta]^\top Q\|\|\|_{\mathsf{op}}\|_{\psi_1} \le \|\|\|P\mathbf{x}^\delta\|\|\|_{\psi_2}\|\|\|Q\mathbf{y}^\delta\|\|\|_{\psi_2} \le (C'')^2\sqrt{M_1M_2}\gamma^{j+k},$$

and since $\mathrm{Off}(A^{j,k}_{t,t'}) = A^{j,k}_{t,t'} - \mathrm{Diag}(A^{j,k}_{t,t'})$, the bound follows from the triangular inequality. □

The following technical lemma will help us to upper-bound σ^2 in Theorem A.1.

Lemma A.3. *Let $\delta_1,\dots,\delta_N$ consists of independent Bernoulli components with probabilities of success $p_1,\dots,p_N$ and set $p_{\min} = \min_{i\le N} p_i$. Let $\mathbf{a},\mathbf{b}\in\mathbb{R}^N$ be two arbitrary vectors. It holds,*

$$\begin{aligned}\mathsf{E}\left(\sum_i \frac{\delta_i}{p_i}a_ib_i\right)^2 \le& p_{\min}^{-1}\|\mathbf{a}\|^2\|\mathbf{b}\|^2,\\ \mathsf{E}\left(\sum_{i\ne j} \frac{\delta_i\delta_j}{p_ip_j}a_ib_j\right)^2 \le& 32p_{\min}^{-2}\|\mathbf{a}\|^2\|\mathbf{b}\|^2 + 4\left(\sum_i a_i\right)^2\left(\sum_i b_i\right)^2.\end{aligned}$$

Additionally, if $\delta'_1,\dots,\delta'_N$ are independent copies of $\delta_1,\dots,\delta_N$, it holds

$$\mathsf{E}\left(\sum_{i,j} \frac{\delta_i\delta'_j}{p_ip_j}a_ib_j\right)^2 \le 4p_{\min}^{-2}\|\mathbf{a}\|^2\|\mathbf{b}\|^2 + 4\left(\sum_i a_i\right)^2\left(\sum_i b_i\right)^2.$$

Proof. It holds,

$$\begin{aligned}\mathsf{E}\left(\sum_i \frac{\delta_i}{p_i}a_ib_i\right)^2 =& \sum_{i,j}\mathsf{E}\frac{\delta_i\delta_j}{p_ip_j}a_ib_ia_jb_j = \sum_{i,j}\{1+\mathbf{1}(i=j)(p_i^{-1}-1)\}a_ib_ia_jb_j\\ \le& \left(\sum_i a_ib_i\right)^2 + (p_{\min}^{-1}-1)\sum_i a_i^2b_i^2\\ \le& \|\mathbf{a}\|^2\|\mathbf{b}\|^2 + (p_{\min}^{-1}-1)\|\mathbf{a}\|^2\|\mathbf{b}\|^2.\end{aligned}$$

To show the second inequality we use decoupling (Theorem 6.1.1 in Vershynin (2018))

and the trivial inequality $(x+y)^2 \le 2x^2 + 2y^2$,

$$\begin{aligned}\mathsf{E}\left(\sum_{i\neq j}\frac{\delta_i\delta_j}{p_ip_j}a_ib_j\right)^2 \le& 2\left(\sum_{i\neq j}a_ib_j\right)^2 + 2\mathsf{E}\left(\sum_{i\neq j}\frac{(\delta_i-p_i)(\delta_j-p_j)}{p_ip_j}a_ib_j\right)^2\\ \le& 2\left(\sum_{i\neq j}a_ib_j\right)^2 + 32\mathsf{E}\left(\sum_{i\neq j}\frac{(\delta_i-p_i)(\delta'_j-p_j)}{p_ip_j}a_ib_j\right)^2.\end{aligned} \tag{A.6}$$

Denote for simplicity $\overline{\delta}_i = \delta_i - p_i$ and $\overline{\delta}'_i = \delta'_i - p_i$. Since the latter are centered we have,

$$\mathsf{E}\left(\sum_{i\neq j}\frac{\overline{\delta}_i\overline{\delta}'_j}{p_ip_j}a_ib_j\right)^2 = \sum_{\substack{i\neq j\\ k\neq l}}\frac{\mathsf{E}\overline{\delta}_i\overline{\delta}_k}{p_ip_k}\frac{\mathsf{E}\overline{\delta}'_j\overline{\delta}'_l}{p_jp_j}a_ia_kb_jb_l \tag{A.7}$$

note that the expectation $\mathsf{E}\overline{\delta}_i\overline{\delta}_k$ is only non-vanishing when $i=k$, in which case it holds $\mathsf{E}\overline{\delta}_i^2 = p_i - p_i^2$. Taking into account similar property of $\mathsf{E}\overline{\delta}'_j\overline{\delta}'_l$ we have that the sum above is equal to

$$\sum_{i\neq j}\frac{(p_i-p_i^2)(p_j-p_j^2)}{p_i^2p_j^2}a_i^2b_j^2 \le (p_{\min}^{-1}-1)^2\sum_{i,j}a_i^2b_j^2 \le (p_{\min}^{-1}-1)^2\|\mathbf{a}\|^2\|\mathbf{b}\|^2.$$

It is left to notice that

$$\left(\sum_{i\neq j}a_ib_j\right)^2 \le 2\left(\sum_{i,j}a_ib_j\right)^2 + 2\left(\sum_i a_ib_j\right)^2 \le 2\left(\sum_i a_i\right)^2\left(\sum_i b_i\right)^2 + 2\|\mathbf{a}\|^2\|\mathbf{b}\|^2,$$

which recalling (A.6) and noting that $32(p_{\min}^{-1}-1)^2+4 \le 32p_{\min}^{-2}$ for $p_{\min}\in[0,1]$, completes the proof.

Similarly to (A.7), we can show the third inequality. □

Now we apply the Bernstein matrix inequality to the sum S_{kj} defined in (A.3), dealing separately with diagonal and off-diagonal parts. After that, we present the proof of Theorem 3.4.

Lemma A.4. *Under the assumptions of Theorem 3.4, it holds for any $u \ge 1$ with probability at least $1-e^{-u}$*

$$\begin{aligned}&\|\!|P\mathrm{diag}\{\mathbf{p}\}^{-1}(\mathrm{Diag}(S_{k,j}) - \mathsf{E}\mathrm{Diag}(S_{k,j}))Q|\!\|_{\mathsf{op}}\\ &\qquad\le C\gamma^{k+j}\left(\sqrt{\frac{M_1\vee M_2(\log N+u)}{Tp_{\min}}}\bigvee\frac{\sqrt{M_1M_2}(\log N+u)}{Tp_{\min}}\right)\end{aligned}$$

where $C = C(K)$ only depends on K.

Proof. Note that,

$$P\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Diag}(S_{kj})Q = T^{-1}\sum_{t=1}^T A_t, \qquad A_t = P\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Diag}(A_{t,t}^{k,j})Q.$$

By Lemma A.2 we have $\|\|\|A_t\|\|_{\mathsf{op}}\|_{\psi_1} \le Cp_{\min}^{-1}\sqrt{M_1M_2}\gamma^{k+j}$. Moreover, using decomposition $Q = \sum_{j=1}^{M_2}\mathbf{u}_j\mathbf{u}_j$, we have

$$\begin{aligned}
\|\|\mathsf{E}A_tA_t^\top\|\|_{\mathsf{op}} \le& \|\|\mathsf{E}\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Diag}(A_{t,t}^{k,j})Q\mathrm{Diag}(A_{t,t}^{k,j})\mathrm{diag}\{\mathbf{p}\}^{-1}\|\|_{\mathsf{op}}\\
\le& \sum_{j=1}^{M_2}\|\|\mathsf{E}\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Diag}(A_{t,t}^{k,j})\mathbf{u}_j\mathbf{u}_j^\top\mathrm{Diag}(A_{t,t}^{k,j})\mathrm{diag}\{\mathbf{p}\}^{-1}\|\|_{\mathsf{op}}\\
\le& \sum_{j=1}^{M_2}\sup_{\|\boldsymbol{\gamma}\|=1}\mathsf{E}(\boldsymbol{\gamma}^\top\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Diag}(A_{t,t}^{k,j})\mathbf{u}_j)^2
\end{aligned}$$

By definition, $\mathrm{Diag}(A_{t,t}^{k,j}) = \mathrm{diag}\{\delta_{ti}x_iy_i\}_{i=1}^N$ for $\mathbf{x} = \Theta^kW_{t-k}$, $\mathbf{y} = \Theta^jW_{t-j}$. Let E_δ denotes the expectation w.r.t. the Bernoulli variables and conditioned on everything else. Setting $\mathbf{a} = (x_1\gamma_1,\dots,x_N\gamma_N)^\top)$ and $\mathbf{b} = (y_1u_1,\dots,y_Nu_N)^\top$, we have by the first inequality of Lemma A.3,

$$\begin{aligned}
\mathsf{E}(\boldsymbol{\gamma}^\top\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Diag}(A_{t,t}^{k,j})\mathbf{u}_j)^2 =& \mathsf{E}\mathsf{E}_\delta\left(\sum_i\gamma_ix_i\frac{\delta_{ti}}{p_i}y_iu_i\right)^2\\
\le& p_{\min}^{-1}\mathsf{E}\|\mathbf{a}\|^2\|\mathbf{b}\|^2\\
\le& p_{\min}^{-1}\mathsf{E}^{1/2}\|\mathbf{a}\|^4\mathsf{E}^{1/4}\|\mathbf{b}\|^4.
\end{aligned}$$

Observe that,

$$\|\mathbf{a}\|^2 = \sum_i\gamma_i^2x_i^2 = \mathbf{x}^\top\mathrm{diag}\{\boldsymbol{\gamma}\}^2\mathbf{x},$$

so since $\mathrm{Tr}(\mathrm{diag}\{\boldsymbol{\gamma}\}^2) = 1$ and due to (A.4) and by Theorem 2.1 Hsu et al. (2012), it holds $\mathsf{E}^{1/2}\|\mathbf{a}\|^4 \le \|\|\mathbf{a}\|^2\|_{\psi_1} \le C'\gamma^{2k}$. Similarly, it holds $\mathsf{E}^{1/2}\|\mathbf{a}\|^4 \le C'\gamma^{2j}$, which together implies

$$\|\|\mathsf{E}A_tA_t^\top\|\|_{\mathsf{op}} \vee \|\|\mathsf{E}A_t^\top A_t^\top\|\|_{\mathsf{op}} \le C''M_2 \vee M_1\gamma^{2k+2j}.$$

Now notice that A_t is not necessary an independent sequence, as A_t depends directly on $(W_{t-k}, W_{t-j}, \boldsymbol{\delta}_t)$, which might intersect with $t' = t + |j-k|$. However, if we take a set $I \subset [1,T]$ such that any two $t,t' \in I$ satisfy $|t'-t| \ne |j-k|$ then the sequence $(A_t)_{t\in I}$ is independent. We separate the whole interval $[1,T]$ into two such independent sets,

$$\begin{aligned}
I_1 =& \{t \in [1,T] :\ \lceil t/|j-k|\rceil \text{ is odd }\},\\
I_2 =& \{t \in [1,T] :\ \lceil t/|j-k|\rceil \text{ is even }\}\\
=& [1,T]\setminus I_1.
\end{aligned} \tag{A.8}$$

Indeed, if for $t,t' \in I_1$ then $\lceil t/|j-k|\rceil$ and $\lceil t'/|j-k|\rceil$ are either equal or differ in at least two, so that in the first case we have $|t-t'| < |j-k|$ and in the second $|t-t'| > |j-k|$. Since both intervals have at most T elements, it holds by Theorem A.1 with probability at least $1-e^{-u}$ for both j,

$$\begin{aligned}
&\|\|\sum_{t\in I_j}A_t - \mathsf{E}A_t\|\|_{\mathsf{op}}\\
&\quad\le C\gamma^{j+k}\left(\sqrt{p_{\min}^{-1}(M_1\vee M_2)T(\log N+u)} \vee p_{\min}^{-1}\sqrt{M_1M_2}(\log N+u)\log T\right),
\end{aligned}$$

so summing up the two and dividing by T, we get the result. □

Lemma A.5. *Under the assumptions of Theorem 3.4, it holds for any $u \geq 1$ with probability at least $1 - e^{-u}$*

$$\begin{aligned}
&\|\!|\!| P\mathrm{diag}\{\mathbf{p}\}^{-1}(\mathrm{Off}(S_{k,j}) - \mathsf{E}\mathrm{Off}(S_{k,j}))\mathrm{diag}\{\mathbf{p}\}^{-1}Q|\!|\!|_{\mathsf{op}} \\
&\qquad \leq C\gamma^{k+j}\left(\sqrt{\frac{M_1 \vee M_2(\log N + u)}{Tp_{\min}^2}} \bigvee \frac{\sqrt{M_1 M_2}(\log N + u)\log T}{Tp_{\min}^2}\right)
\end{aligned}$$

where $C = C(K)$ only depends on K.

Proof. It holds,

$$\begin{aligned}
P\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Off}(S_{kj})\mathrm{diag}\{\mathbf{p}\}^{-1}Q &= T^{-1}\sum_{t=1}^{T} B_t, \\
B_t &= P\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Off}(A_{t,t}^{k,j})\mathrm{diag}\{\mathbf{p}\}^{-1}Q.
\end{aligned}$$

By Lemma A.2 we have $\|\|\!|B_t|\!|\!|_{\mathsf{op}}\|_{\psi_1} \leq Cp_{\min}^{-2}\sqrt{M_1M_2}\gamma^{k+j}$. Using decomposition $Q = \sum_{j=1}^{M_2}\mathbf{u}_j\mathbf{u}_j$ with $\|\mathbf{u}_j\| = 1$ we get that

$$\begin{aligned}
\|\!|\!|\mathsf{E}B_tB_t^\top|\!|\!|_{\mathsf{op}} \leq& \|\!|\!|\mathsf{E}\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Off}(A_{t,t}^{k,j})\mathrm{diag}\{\mathbf{p}\}^{-1}Q\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Off}(A_{t,t}^{k,j})\mathrm{diag}\{\mathbf{p}\}^{-1}|\!|\!|_{\mathsf{op}} \\
\leq& \sum_{j=1}^{M_2}\|\!|\!|\mathsf{E}\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Off}(A_{t,t}^{k,j})\mathrm{diag}\{\mathbf{p}\}^{-1}\mathbf{u}_j\mathbf{u}_j^\top\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Off}(A_{t,t}^{k,j})\mathrm{diag}\{\mathbf{p}\}^{-1}|\!|\!|_{\mathsf{op}} \\
\leq& \sum_{j=1}^{M_2}\sup_{\|\boldsymbol{\gamma}\|=1}\mathsf{E}(\boldsymbol{\gamma}^\top\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Off}(A_{t,t}^{k,j})\mathrm{diag}\{\mathbf{p}\}^{-1}\mathbf{u}_j)^2
\end{aligned}$$

Again, using the notation $\mathbf{x} = \Theta^k W_{t-k}$, $\mathbf{y} = \Theta^j W_{t-j}$ and $\mathbf{a} = \mathrm{diag}\{\boldsymbol{\gamma}\}\mathbf{x}$, $\mathbf{b} = \mathrm{diag}\{\mathbf{u}\}\mathbf{y}$, we have $\mathrm{Off}(A_{t,t}^{j,k}) = \mathrm{Off}(\mathbf{x}\mathbf{y}^\top)$. Therefore, by Lemma A.3

$$\begin{aligned}
\mathsf{E}(\boldsymbol{\gamma}^\top\mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Off}(A_{t,t}^{k,j})\mathrm{diag}\{\mathbf{p}\}^{-1}\mathbf{u}_j)^2 =& \mathsf{E}\mathsf{E}_\delta\left(\sum_{i\neq j}\gamma_i\frac{\delta_{it}}{p_i}x_iy_j\frac{\delta_{jt}}{\delta_j}u_j\right)^2 \\
=& \mathsf{E}\mathsf{E}_\delta\left(\sum_{i\neq j}\frac{\delta_{it}}{p_i}\frac{\delta_{jt}}{\delta_j}a_ib_j\right)^2 \\
\leq& 32p_{\min}^{-2}\mathsf{E}\|\mathbf{a}\|^2\|\mathbf{b}\|^2 + 4\mathsf{E}\left(\sum_i a_i\right)^2\left(\sum_i b_i\right)^2.
\end{aligned}$$

From the proof of Lemma A.5 we know that $\mathsf{E}\|\mathbf{a}\|^2\|\mathbf{b}\|^2 \leq C'\gamma^{2k+2j}$. Moreover, we have $\sum_i a_i = \boldsymbol{\gamma}^\top\mathbf{x}$ and $\sum_i b_i = \mathbf{u}^\top\mathbf{y}$. Thus, by (A.5) it holds $\mathsf{E}^{1/4}\|\boldsymbol{\gamma}^\top\mathbf{x}\|^4 \leq \|\boldsymbol{\gamma}^\top\mathbf{x}\|_{\psi_2} \leq C'\gamma^j$ and, similarly, $\mathsf{E}^{1/4}\|\mathbf{u}^\top\mathbf{y}\|^4 \leq C'\gamma^k$. Putting those bounds together and applying Cauchy-Schwarz inequality, we have

$$\|\!|\!|\mathsf{E}B_tB_t^\top|\!|\!|_{\mathsf{op}} \leq C''p_{\min}^{-2}M_2\gamma^{2k+2j}.$$

By analogy,

$$\|\!|\mathsf{E}B_tB_t^\top|\!\|_{\mathsf{op}} \vee \|\!|\mathsf{E}B_t^\top B_t|\!\|_{\mathsf{op}} \le C'' p_{\min}^{-2} M_1 \vee M_2 \gamma^{2k+2j}.$$

Applying the same sample splitting (A.8) we obtain the bound

$$\|\!|\sum_t A_t - \mathsf{E}A_t|\!\|_{\mathsf{op}} \le C\gamma^{j+k}\left(\sqrt{p_{\min}^{-2}(M_1 \vee M_2)T(\log N + u)} \vee p_{\min}^{-2}\sqrt{M_1M_2}(\log N + u)\right),$$

which divided by T provides the result. □

Proof of Theorem 3.4. Set,

$$S_{k,j}^{\boldsymbol{\delta}} = \mathrm{diag}\{\mathbf{p}\}^{-1}\mathrm{Diag}(S_{k,j}) - \mathrm{diag}\{\boldsymbol{\delta}\}^{-1}\mathrm{Off}(S_{k,j})\mathrm{diag}\{\boldsymbol{\delta}\}^{-1},$$

so that by the union of bounds in Lemmas A.5, A.4 for any $u \ge 1$

$$\|\!|P(S_{k,j}^{\boldsymbol{\delta}} - \mathsf{E}S_{k,j}^{\boldsymbol{\delta}})Q|\!\|_{\mathsf{op}} > C\gamma^{k+j}\left(\sqrt{\frac{M_1 \vee M_2(\log N + u)}{Tp_{\min}^2}} \bigvee \frac{\sqrt{M_1M_2}(\log N + u)}{Tp_{\min}^2}\right)$$

holds with probability at least $1 - e^{-u}$. Take a union of those bounds for every k, j with $u = u_{k,j} = k + j + 1 + u'$. The total probability of complementary event is at most

$$\sum_{k,j\ge 0} e^{-k-j-1-u} = e^{-1-u}\left(\sum_{k\ge 0} e^{-k}\right)^2 = e^{-u}/(e-1) < e^{-u}.$$

On such event it holds

$$\begin{aligned}
\|\!|P(\hat\Sigma - \mathsf{E}\Sigma)Q|\!\|_{\mathsf{op}} \le& \sum_{k,j\ge 0} \|\!|P(S_{k,j}^{\boldsymbol{\delta}} - \mathsf{E}S_{k,j}^{\boldsymbol{\delta}})Q|\!\|_{\mathsf{op}} \\
\le& C\sum_{k,j\ge 0}\gamma^{k+j}\left(\sqrt{\frac{M_1 \vee M_2(\log N + u_{k,j})}{Tp_{\min}^2}} \bigvee \frac{\sqrt{M_1M_2}(\log N + u_{k,j})}{Tp_{\min}^2}\right) \\
\le& C'\left[\sum_{k,j\ge 0}\gamma^{k+j}\right]\left(\sqrt{\frac{(M_1 \vee M_2)\log N}{Tp_{\min}^2}} \bigvee \frac{\sqrt{M_1M_2}\log N}{Tp_{\min}^2}\right) \\
&+ C\left[\sum_{k,j}(k+j)\gamma^{k+j}\right]\left(\sqrt{\frac{(M_1 \vee M_2)u}{Tp_{\min}^2}} \bigvee \frac{\sqrt{M_1M_2}u}{Tp_{\min}^2}\right),
\end{aligned}$$

which completes the proof due to the equalities

$$\begin{aligned}
\sum_{k,j\ge 0}\gamma^{k+j} &= \left(\sum_{k\ge 0}\gamma^k\right)^2 = \frac{1}{(1-\gamma)^2} \\
\sum_{k,j\ge 0}(k+j)\gamma^{k+j} &= 2\sum_{k,j\ge 0} k\gamma^{k+j} = \frac{2}{(1-\gamma)}\sum_{k\ge 0} k\gamma^k = \frac{2}{(1-\gamma)^3}.
\end{aligned}$$

□

Proof of Theorem 3.5. Recall the definition,

$$A^{k,j}_{t,t'} = \mathrm{diag}\{\boldsymbol{\delta}_t\}\Theta^k W_{t-k} W^\top_{t'-j}[\Theta^j]^\top \mathrm{diag}\{\boldsymbol{\delta}_{t'}\}.$$

Then, it holds

$$Z_t Z^\top_{t+1} = \sum_{k,j\geq 0} \mathrm{diag}\{\boldsymbol{\delta}_t\}\Theta^k W_{t-k} W^\top_{t+1-j}[\Theta^j]^\top \mathrm{diag}\{\boldsymbol{\delta}_{t+1}\} = \sum_{k,j\geq 0} A^{k,j}_{t,t+1},$$

and the decomposition takes place,

$$A^* = \sum_{k,j\geq 0} S_{k,j}, \qquad S_{k,j} = \frac{1}{T-1}\sum_{t=1}^{T-1} A^{k,j}_{t,t+1}.$$

We first apply the Bernstein matrix inequality for each $S_{k,j}$ separately. Observe that

$$P\mathrm{diag}\{\mathbf{p}\}^{-1} S_{k,j} \mathrm{diag}\{\mathbf{p}\}^{-1} Q = \frac{1}{T-1}\sum_{t=1}^{T-1} B_t, \qquad B_t = P\mathrm{diag}\{\mathbf{p}\}^{-1} A^{k,j}_{t,t+1} \mathrm{diag}\{\mathbf{p}\}^{-1} Q.$$

By Lemma A.2 each term satisfies

$$\max_t \| |\!|\!| B_t |\!|\!|_{\mathsf{op}} \|_{\psi_1} \leq C\sqrt{M_1 M_2}\gamma^{k+j}.$$

Furthermore, let $Q = \sum_{j=1}^{M_2} \mathbf{u}_j \mathbf{u}_j^\top$ with unit vectors $\mathbf{u}_j$. Also, denoting $\mathbf{x} = \Theta^k W_{t-k}$ and $\mathbf{y} = \Theta^k W_{t+1-k}$ it holds $A^{k,j}_{t,t+1} = \mathrm{diag}\{\boldsymbol{\delta}_t\}\mathbf{x}\mathbf{y}^\top \mathrm{diag}\{\boldsymbol{\delta}_{t+1}\}$. Then, using Lemma A.3 we have for any unit $\boldsymbol{\gamma} \in \mathbb{R}^N$,

$$\begin{aligned}
&\mathsf{E}(\boldsymbol{\gamma}^\top \mathrm{diag}\{\mathbf{p}\}^{-1} A^{k,j}_{t,t+1} \mathrm{diag}\{\mathbf{p}\}^{-1}\mathbf{u}_j)^2 \\
&\quad = \mathsf{E}\mathsf{E}_\delta \left(\sum_{i,j} \gamma_i x_i \frac{\delta_{ti}}{p_i}\frac{\delta_{t+1,j}}{p_j} y_j u_j \right)^2 \\
&\quad \leq p_{\min}^{-2}\mathsf{E}\|\mathrm{diag}\{\boldsymbol{\gamma}\}\mathbf{x}\|^2 \|\mathrm{diag}\{\mathbf{u}\}\mathbf{y}\|^2 + \mathsf{E}(\boldsymbol{\gamma}^\top\mathbf{x})(\mathbf{u}^\top\mathbf{y})^2,
\end{aligned}$$

which due to the subgaussianity of $\mathbf{x}$ and $\mathbf{y}$ yields,

$$\begin{aligned}
\mathsf{E}\|\mathrm{diag}\{\boldsymbol{\gamma}\}\mathbf{x}\|^2 \|\mathrm{diag}\{\mathbf{u}\}\mathbf{y}\|^2 &\leq \mathsf{E}^{1/2}\|\mathrm{diag}\{\boldsymbol{\gamma}\}\mathbf{x}\|^4 \mathsf{E}^{1/2}\|\mathrm{diag}\{\mathbf{u}\}\mathbf{y}\|^4 \\
&\leq C'\gamma^{2k+2j} \\
\mathsf{E}(\boldsymbol{\gamma}^\top\mathbf{x})(\mathbf{u}^\top\mathbf{y})^2 &\leq \mathsf{E}^{1/2}(\boldsymbol{\gamma}^\top\mathbf{x})^4 \mathsf{E}^{1/2}(\mathbf{u}^\top\mathbf{y})^4 \\
&\leq C'\gamma^{2k+2j}.
\end{aligned}$$

Therefore, we get that

$$| \! | \! | \mathsf{E} B_t B_t^\top | \! | \! |_{\mathsf{op}} = \sup_{\|\boldsymbol{\gamma}\|=1} \sum_{j=1}^{M_2} \mathsf{E}\left(\boldsymbol{\gamma}^\top \mathrm{diag}\{\mathbf{p}\}^{-1} A^{k,j}_{t,t+1} \mathrm{diag}\{\mathbf{p}\}^{-1}\mathbf{u}_j\right)^2 \leq C'' p_{\min}^{-2} M_2 \gamma^{2k+2j}.$$

Using similar derivations we can arrive at

$$\sigma^2 = \|\!|\mathsf{E} B_t B_t^\top |\!\|_{\mathsf{op}} \vee \|\!|\mathsf{E} B_t^\top B_t |\!\|_{\mathsf{op}} \le C'' p_{\min}^{-2}(M_1 \vee M_2)\gamma^{2k+2j}.$$

Now we separate the indices $t = 1, \dots, T$ into four subsets, such that each corresponds to a set of independent matrices B_t. Since each B_t is generated by $W_{t-k}, W_{t+1-j}, \boldsymbol{\delta}_t$, and $\boldsymbol{\delta}_{t+1}$, we need to ensure that none of the pair of indices t, t' from the same subset satisfies $|t - t'| = |k - j + 1|$ nor $|t - t'| = 1$. It can be satisfied by the following partition. First, we split the indices into two subsets with odd and even indices, respectively, so that none of the subsets contains two indices with $|t - t'| = 1$. Then, both of the subsets need to be separated into two according to the scheme (A.8), so that the assertion $|t - t'| = |k - j + 1|$ is avoided within each subset. Therefore, applying the Bernstein inequality, Theorem A.1, to each sum separately and summing them up, we get that for any $u \ge 1$ with probability at least $1 - e^{-u}$,

$$\begin{aligned}\|\!|P\mathrm{diag}\{\boldsymbol{\delta}\}^{-1}(S_{k,j} - \mathsf{E}S_{k,j})\mathrm{diag}\{\boldsymbol{\delta}\}^{-1}Q|\!\|_{\mathsf{op}} \\ \le C\left(\sqrt{p_{\min}^{-2}(M_1 \vee M_2)T(\log N + u)} \bigvee \sqrt{M_1 M_2}(\log N + u)\log T\right).\end{aligned}$$

Similarly to the proof of Theorem 3.4, we take the union of those bounds for every i, j with $u = j + k + u'$ and then the result follows. □

B LASSO and missing observations

Suppose, we observe a signal $\mathbf{y} \in \mathbb{R}^n$ of the form

$$\mathbf{y} = \Phi\mathbf{b}^* + \boldsymbol{\varepsilon},$$

where $\Phi = [\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_p] \in \mathbb{R}^{n \times p}$ is a dictionary of words $\boldsymbol{\phi}_j \in \mathbb{R}^n$ and $\mathbf{b}^*$ is some sparse parameter with support $\Lambda \subset \{1, \dots, p\}$. We want to recover the exact sparse representation by solving a quadratic program

$$\frac{1}{2}\|\mathbf{y} - \Phi\mathbf{b}\|^2 + \gamma\|\mathbf{b}\|_1 \to \min_{\mathbf{b} \in \mathbb{R}^p}. \tag{B.1}$$

Denote by $\mathbb{R}^\Lambda$ the set of vectors with elements indexed by Λ, for $\mathbf{b} \in \mathbb{R}^n$ let $\mathbf{x}_\Lambda \in \mathbb{R}^\Lambda$ be the result of taking only elements indexed by Λ. With some abuse of notation we will associate every vector $\mathbf{x}_\Lambda \in \mathbb{R}^\Lambda$ with a vector $\mathbf{x}$ from $\mathbb{R}^n$ that has same coefficients on Λ and zeros elsewhere. Let $\Phi_\Lambda = [\boldsymbol{\phi}_j]_{j \in \Lambda}$ be a subdictionary composed of words indexed by Λ, and P_Λ is the projector onto the corresponding subspace.

The following sufficient conditions for the global minimizer of (B.1) to be supported on Λ are due to Tropp (2006), who uses the notion of *exact recovery coefficient*,

$$\mathrm{ERC}_\Phi(\Lambda) = 1 - \max_{j \notin \Lambda} \|\Phi_\Lambda^+ \boldsymbol{\phi}_j\|_1,$$

The results are summarized in the next theorem.

Theorem B.1 (Tropp (2006))**.** *Let* $\tilde{\mathbf{b}}$ *be a solution to* (B.1)*. Suppose that* $\|\Phi^\top\boldsymbol{\varepsilon}\|_\infty \le \gamma\mathrm{ERC}(\Lambda)$*. Then,*

- *the support of* $\tilde{\mathbf{b}}$ *is contained in* Λ*;*
- *the distance between* $\tilde{\mathbf{b}}$ *and optimal (non-penalized) parameter satisfies,*

$$\|\tilde{\mathbf{b}} - \mathbf{b}^*\|_\infty \le \|\Phi_\Lambda^+ \boldsymbol{\varepsilon}\|_\infty + \gamma \|(\Phi_\Lambda \Phi_\Lambda^\top)^{-1}\|_{1,\infty},$$
$$\|\Phi_\Lambda(\tilde{\mathbf{b}} - \mathbf{b}^*) - P_\Lambda \boldsymbol{\varepsilon}\|_2 \le \gamma \|(\Phi_\Lambda^+)^\top\|_{2,\infty};$$

In what follows, we want to extend this result for the possibility of using missing observations model. Observe that the program (B.1) is equivalent to

$$\frac{1}{2}\mathbf{b}^\top[\Phi^\top\Phi]\mathbf{b} - \mathbf{b}^\top[\Phi^\top \mathbf{y}] + \gamma\|\mathbf{b}\|_1 \to \min_{\mathbf{b}\in\mathbb{R}^p},$$

so that the minimization procedure only depends on $D = \Phi^\top\Phi$ and $\mathbf{c} = \Phi^\top\mathbf{y}$. Suppose that instead we have only the access to some estimators $\hat{D} \ge 0$ and $\hat{\mathbf{c}}$ that are close enough to the original matrix and vector, respectively, which may come e.g., from missing observations model. Then, we can solve instead the following problem,

$$\frac{1}{2}\mathbf{b}^\top \hat{D}\mathbf{b} - \mathbf{b}^\top \hat{\mathbf{c}} + \gamma\|\mathbf{b}\|_1 \to \min_{\mathbf{b}\in\mathbb{R}^p}. \tag{B.2}$$

In what follows, we provide a slight extension of Tropp's result towards missing observations, the proof mainly follows the same steps.

Below, for a matrix D and two sets of indices A, B, we denote the submatrix on those indices as $D_{A,B}$, and for a vector $\mathbf{c}$, the corresponding subvector is $\mathbf{c}_A$.

Lemma B.2. *Suppose that*

$$\|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\hat{\mathbf{c}}_\Lambda - \hat{\mathbf{c}}_{\Lambda^c}\|_\infty \le \gamma(1 - \|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\|_{1,\infty}).$$

Then, the solution $\tilde{\mathbf{b}}$ *to* (B.2) *is supported on* Λ*.*

Proof. Let $\tilde{\mathbf{b}}$ be the solution to (B.2) with the restriction $\mathrm{supp}(\mathbf{b}) \subset \Lambda$. Since $\hat{D} \ge 0$ this is a convex problem and therefore the solution is unique and satisfies

$$\hat{D}_{\Lambda,\Lambda}\tilde{\mathbf{b}} - \hat{\mathbf{c}}_\Lambda + \gamma\mathbf{g} = 0, \qquad \mathbf{g} \in \partial\|\tilde{\mathbf{b}}\|_1,$$

where $\partial f(\mathbf{b})$ denotes the subdifferential of a convex function f at a point $\mathbf{b}$, in the case of ℓ_1 norm we have $\|\mathbf{g}\|_\infty \le 1$. Thus,

$$\tilde{\mathbf{b}} = \hat{D}^{-1}_{\Lambda,\Lambda}\hat{\mathbf{c}}_\Lambda - \gamma\hat{D}^{-1}_{\Lambda,\Lambda}\mathbf{g}. \tag{B.3}$$

Next, we want to check that $\tilde{\mathbf{b}}$ is a global minimizer. To do so, let us compare the objective function at a point $\overline{\mathbf{b}} = \tilde{\mathbf{b}} + \delta\mathbf{e}_j$ for arbitrary index $j \notin \Lambda$. Since $\|\overline{\mathbf{b}}\|_1 = \|\tilde{\mathbf{b}}\|_1 + |\delta|$, we have

$$\begin{aligned} L(\tilde{\mathbf{b}}) - L(\overline{\mathbf{b}}) &= \frac{1}{2}\tilde{\mathbf{b}}^\top\hat{D}\tilde{\mathbf{b}} - \frac{1}{2}\overline{\mathbf{b}}^\top\hat{D}\overline{\mathbf{b}} - \hat{\mathbf{c}}^\top(\tilde{\mathbf{b}} - \overline{\mathbf{b}}) - \gamma|\delta| \\ &= \frac{\delta^2}{2}\mathbf{e}_j^\top\hat{D}\mathbf{e}_j + |\delta|\gamma - \delta\mathbf{e}_j^\top\hat{D}\tilde{\mathbf{b}} + \delta\widehat{c}_j \\ &> |\delta|\gamma - \delta\mathbf{e}_j^\top\hat{D}\tilde{\mathbf{b}} + \delta\widehat{c}_j, \end{aligned}$$

where the latter comes from the fact that $\hat{D}$ is positively definite. Applying the equality (B.3) yields,

$$\mathbf{e}_j^\top \hat{D}\tilde{\mathbf{b}} = \hat{D}_{j,\Lambda}\hat{D}_{\Lambda,\Lambda}^{-1}\hat{\mathbf{c}}_\Lambda - \gamma \hat{D}_{j,\Lambda}\hat{D}_{\Lambda,\Lambda}^{-1}\mathbf{g},$$

therefore, taking into account $\|\mathbf{g}\|_\infty \le 1$ we have,

$$L(\tilde{\mathbf{b}}) - L(\overline{\mathbf{b}}) > |\delta| \left[\gamma(1 - \|\hat{D}_{\Lambda^c,\Lambda}\hat{D}_{\Lambda,\Lambda}^{-1}\|_{1,\infty}) - \left|\hat{D}_{j,\Lambda}\hat{D}_{\Lambda,\Lambda}^{-1}\hat{\mathbf{c}}_\Lambda - \widehat{c}_j\right|\right],$$

where the right-hand side is nonnegative by the condition of the lemma. Since $j \notin \Lambda$ is arbitrary, $\tilde{\mathbf{b}}$ is a global solution as well.

□

Remark B.1. It is not hard to see that in the exact case $\hat{D} = \Phi^\top\Phi$ and $\hat{\mathbf{c}} = \Phi^\top\mathbf{y}$ the condition of the lemma above turns into the condition $\|\Phi_{\Lambda^c}^\top P_\Lambda \boldsymbol{\varepsilon}\|_\infty \le \gamma \mathrm{ERC}(\Lambda)$ of Theorem B.1.

Since we are particularly interested in applications to time series, the features matrix Φ should in fact be random, thus stating a ERC-like condition onto it might result in additional unnecessary technical difficulties. Instead, let us assume that there is some other matrix $\bar{D}$, potentially the expectation of $\Phi^\top\Phi$, such that it is close enough to $\hat{D}$ (with some probability, but we are stating all the results deterministically in this section), and the value that controls the exact recovery looks like

$$\mathrm{ERC}(\Lambda; \bar{D}) = 1 - \|\bar{D}_{\Lambda^c,\Lambda}\bar{D}_{\Lambda,\Lambda}^{-1}\|_{1,\infty}.$$

Additionally, we set $\bar{\mathbf{c}} = \bar{D}\mathbf{b}^* = \bar{D}_{\cdot,\Lambda}\mathbf{b}_\Lambda^*$ — the vector that $\hat{\mathbf{c}}$ is intended to approximate. Note that in this case we have $\bar{D}_{\Lambda^c,\Lambda}\bar{D}_{\Lambda,\Lambda}^{-1}\bar{\mathbf{c}}_\Lambda - \bar{\mathbf{c}}_{\Lambda^c} = \bar{D}_{\Lambda^c,\Lambda}\mathbf{b}_\Lambda^* - \bar{\mathbf{c}}_{\Lambda^c} = 0$, thus the conditions of Lemma B.2 hold for $\bar{D}, \bar{\mathbf{c}}$ once $\mathrm{ERC}(\Lambda; \bar{D})$ and γ are nonnegative. In what follows, we control the values appearing in the lemma for $\hat{D}$ and $\hat{\mathbf{c}}$ through the differences between $\bar{\mathbf{c}}$, $\bar{D}$ and $\hat{\mathbf{c}}$, $\hat{D}$, respectively, thus allowing the exact recovery of the sparsity pattern. Lemma 7.7

Corollary B.3. *Let $\bar{D}$ and $\bar{\mathbf{c}}$ be such that $\bar{\mathbf{c}} = \bar{D}\mathbf{b}^*$. Assume that*

$$\|\hat{\mathbf{c}} - \bar{\mathbf{c}}\|_\infty \le \delta_c, \qquad \|\bar{D}_{\Lambda,\Lambda}^{-1}(\hat{\mathbf{c}}_\Lambda - \bar{\mathbf{c}}_\Lambda)\|_\infty \le \delta_c', \qquad \|\bar{D}_{\Lambda,\Lambda}^{-1}(\hat{D}_{\Lambda,\cdot} - \bar{D}_{\Lambda,\cdot})\|_{\infty,\infty} \le \delta_D,$$
$$\|(\hat{D}_{\cdot,\Lambda} - \bar{D}_{\cdot,\Lambda})\mathbf{b}_\Lambda^*\|_\infty \le \delta_D', \qquad \|\bar{D}_{\Lambda,\Lambda}^{-1}(\bar{D}_{\Lambda,\Lambda} - \hat{D}_{\Lambda,\Lambda})\mathbf{b}_\Lambda^*\|_\infty \le \delta_D''.$$

Suppose, $\mathrm{ERC}(\Lambda) \ge 3/4$ *and*

$$3\delta_c + 3\delta_D' \le \gamma, \qquad s\delta_D \le \frac{1}{16},$$

where $|\Lambda| = s$. Then, the solution to (B.2) *is supported on a subset of Λ and satisfies*

$$\tilde{\mathbf{b}}_\Lambda = \hat{D}_{\Lambda,\Lambda}^{-1}\hat{\mathbf{c}}_\Lambda - \gamma\hat{D}_{\Lambda,\Lambda}^{-1}\mathbf{g}, \tag{B.4}$$

with some $\mathbf{g} \in \mathbb{R}^s$ satisfying $\|\mathbf{g}_\Lambda\|_\infty \le 1$ and the max-norm error satisfies

$$\|\tilde{\mathbf{b}} - \mathbf{b}^*\|_\infty \le 2(\delta_D'' + \delta_c' + \gamma\|\bar{D}_{\Lambda,\Lambda}^{-1}\|_{1,\infty}),$$

while the ℓ_2-norm error satisfies

$$\|\tilde{\mathbf{b}} - \mathbf{b}^*\| \le 2\sqrt{s}(\delta''_D + \delta'_c + \gamma\sigma_{\min}^{-1}).$$

If additionally $2(\delta''_D+\delta'_c+\gamma\|\bar{D}^{-1}_{\Lambda,\Lambda}\|_{1,\infty}) \le \min_{j\in\Lambda}|\mathbf{b}^*_j|$, *then we have the exact recovery, so that the following equality takes place*

$$\tilde{\mathbf{b}}_\Lambda = \hat{D}^{-1}_{\Lambda,\Lambda}\hat{\mathbf{c}}_\lambda - \gamma\hat{D}^{-1}_{\Lambda,\Lambda}\mathbf{s}_\Lambda,$$

where $\mathbf{s} = sign(\mathbf{b}^*)$.

Proof. First, observe that $D_{\Lambda^c,\Lambda}D^{-1}_{\Lambda,\Lambda}\mathbf{c}_\Lambda - \mathbf{c}_{\Lambda^c} = \Phi^\top_{\Lambda^c}(\Phi^+_\Lambda\mathbf{y} - \mathbf{y}) = \Phi^\top_{\Lambda^c}(P_\Lambda - I)\boldsymbol{\varepsilon}$. By Lemma B.4 we have,

$$\|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\|_{1,\infty} \le \|\bar{D}_{\Lambda^c,\Lambda}\bar{D}^{-1}_{\Lambda,\Lambda}\|_{1,\infty} + 4s\delta_D \le 1/2,$$

while since $\bar{\mathbf{c}}_{\Lambda^c} = \bar{D}_{\Lambda^c,\Lambda}\mathbf{b}^*_\Lambda = \bar{D}_{\Lambda^c,\Lambda}\bar{D}^{-1}_{\Lambda,\Lambda}\bar{\mathbf{c}}_\Lambda$,

$$\begin{aligned}
\|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\hat{\mathbf{c}}_\Lambda - \hat{\mathbf{c}}_{\Lambda^c}\|_\infty &\le \|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\hat{\mathbf{c}}_\Lambda - \bar{D}_{\Lambda^c,\Lambda}\bar{D}^{-1}_{\Lambda,\Lambda}\bar{\mathbf{c}}_\Lambda\|_\infty + \|\hat{\mathbf{c}}_{\Lambda^c} - \bar{\mathbf{c}}_{\Lambda^c}\|_\infty \\
&\le \|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}(\hat{\mathbf{c}}_\Lambda - \bar{\mathbf{c}}_\Lambda)\|_\infty + \|\hat{D}_{\Lambda^c,\Lambda}(\hat{D}^{-1}_{\Lambda,\Lambda} - \bar{D}^{-1}_{\Lambda,\Lambda})\bar{\mathbf{c}}_\Lambda\|_\infty \\
&\quad + \|(\hat{D}_{\Lambda^c,\Lambda} - \bar{D}_{\Lambda^c,\Lambda})\bar{D}^{-1}_{\Lambda,\Lambda}\bar{\mathbf{c}}_\Lambda\|_\infty + \delta_c \\
&\le \|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}(\hat{\mathbf{c}}_\Lambda - \bar{\mathbf{c}}_\Lambda)\|_\infty + \|\hat{D}_{\Lambda^c,\Lambda}(\hat{D}^{-1}_{\Lambda,\Lambda} - \bar{D}^{-1}_{\Lambda,\Lambda})\bar{\mathbf{c}}_\Lambda\|_\infty + \delta'_D + \delta_c.
\end{aligned}$$

Here, $\|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}(\hat{\mathbf{c}}_\Lambda - \bar{\mathbf{c}}_\Lambda)\|_\infty \le \delta_c/2$ due to $\|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\|_{1,\infty} \le 1/2$. Moreover, we have

$$\begin{aligned}
\|\hat{D}_{\Lambda^c,\Lambda}(\hat{D}^{-1}_{\Lambda,\Lambda} - \bar{D}^{-1}_{\Lambda,\Lambda})\bar{\mathbf{c}}_\Lambda\|_\infty &= \|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}(\bar{D}_{\Lambda,\Lambda} - \hat{D}_{\Lambda,\Lambda})\bar{D}^{-1}_{\Lambda,\Lambda}\bar{\mathbf{c}}_\Lambda\|_\infty \\
&\le \|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\|_{1,\infty}\|(\bar{D}_{\Lambda,\Lambda} - \hat{D}_{\Lambda,\Lambda})\bar{D}^{-1}_{\Lambda,\Lambda}\bar{\mathbf{c}}_\Lambda\|_\infty \\
&\le \delta'_D/2.
\end{aligned}$$

Using the condition on γ, we get that

$$\|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\hat{\mathbf{c}}_\Lambda - \hat{\mathbf{c}}_{\Lambda^c}\|_\infty \le \frac{3}{2}(\delta'_D + \delta_c) \le \frac{\gamma}{2} \le \gamma(1 - \|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\|_{1,\infty}),$$

so that the conditions of Lemma B.2 are satisfied and (B.4) takes place. Therefore, we can write

$$\begin{aligned}
\tilde{\mathbf{b}}_\Lambda - \mathbf{b}^*_\Lambda &= \hat{D}^{-1}_{\Lambda,\Lambda}\hat{\mathbf{c}}_\Lambda - \bar{D}^{-1}_{\Lambda,\Lambda}\bar{\mathbf{c}}_\Lambda - \gamma\hat{D}^{-1}_{\Lambda,\Lambda}\mathbf{g}, \\
&= \hat{D}^{-1}_{\Lambda,\Lambda}(\bar{D}_{\Lambda,\Lambda} - \hat{D}_{\Lambda,\Lambda})\bar{D}^{-1}_{\Lambda,\Lambda}\bar{\mathbf{c}}_\Lambda + \hat{D}^{-1}_{\Lambda,\Lambda}(\hat{\mathbf{c}}_\Lambda - \bar{\mathbf{c}}_\Lambda) - \gamma\hat{D}^{-1}_{\Lambda,\Lambda}\mathbf{g} \\
&= \hat{D}^{-1}_{\Lambda,\Lambda}(\bar{D}_{\Lambda,\Lambda} - \hat{D}_{\Lambda,\Lambda})\mathbf{b}^*_\Lambda + \hat{D}^{-1}_{\Lambda,\Lambda}(\hat{\mathbf{c}}_\Lambda - \bar{\mathbf{c}}_\Lambda) - \gamma\hat{D}^{-1}_{\Lambda,\Lambda}\mathbf{g} \\
&= \hat{D}^{-1}_{\Lambda,\Lambda}\bar{D}_{\Lambda,\Lambda}\left(\bar{D}^{-1}_{\Lambda,\Lambda}(\bar{D}_{\Lambda,\Lambda} - \hat{D}_{\Lambda,\Lambda})\mathbf{b}^*_\Lambda + \bar{D}^{-1}_{\Lambda,\Lambda}(\hat{\mathbf{c}}_\Lambda - \bar{\mathbf{c}}_\Lambda) - \gamma\bar{D}^{-1}_{\Lambda,\Lambda}\mathbf{g}\right)
\end{aligned}$$

By Lemma B.4 we have $\|\hat{D}^{-1}_{\Lambda,\Lambda}\bar{D}_{\Lambda,\Lambda}\|_{\infty\mapsto\infty} \le 2$ so that

$$\|\tilde{\mathbf{b}}_\Lambda - \mathbf{b}^*_\Lambda\|_\infty \le 2\|\bar{D}^{-1}_{\Lambda,\Lambda}(\bar{D}_{\Lambda,\Lambda} - \hat{D}_{\Lambda,\Lambda})\mathbf{b}^*_\Lambda\|_\infty + 2\|\bar{D}^{-1}_{\Lambda,\Lambda}(\hat{\mathbf{c}}_\Lambda - \bar{\mathbf{c}}_\Lambda)\|_\infty + 2\gamma\|\bar{D}^{-1}_{\Lambda,\Lambda}\|_{1,\infty}.$$

Since we also have $\|\!|\hat{D}^{-1}_{\Lambda,\Lambda}\bar{D}_{\Lambda,\Lambda}|\!\|_{\mathsf{op}} \le 2$ and $\|\mathbf{g}\| \le \sqrt{s}$, it holds

$$\|\tilde{\mathbf{b}}_\Lambda - \mathbf{b}^*_\Lambda\| \le 2\sqrt{s}\left(\|\bar{D}^{-1}_{\Lambda,\Lambda}(\bar{D}_{\Lambda,\Lambda} - \hat{D}_{\Lambda,\Lambda})\mathbf{b}^*_\Lambda\|_\infty + \|\bar{D}^{-1}_{\Lambda,\Lambda}(\hat{\mathbf{c}}_\Lambda - \bar{\mathbf{c}}_\Lambda)\|_\infty + \gamma|\!|\!|\bar{D}^{-1}_{\Lambda,\Lambda}|\!|\!|_{\mathsf{op}}\right).$$

□

Before we proceed with the proof of this corollary, we present a technical lemma that collects some trivial inequalities.

Lemma B.4. *Set* $\delta_c = \|\hat{\mathbf{c}} - \bar{\mathbf{c}}\|_\infty$, $\delta_D = \|(\hat{D}_{\Lambda^c,\Lambda} - \bar{D}_{\Lambda^c,\Lambda})\bar{D}^{-1}_{\Lambda,\Lambda}\|_{\infty,\infty}$. *Suppose,* $\|\bar{D}_{\Lambda^c\Lambda}\bar{D}^{-1}_{\Lambda\Lambda}\|_{1,\infty} \le 1$ *and* $s\delta_D \le 1/2$. *It holds,*

- *for any* $q \ge 1$

$$\|D_{\Lambda,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\|_{q\to q} \le 2, \qquad \|\hat{D}^{-1}_{\Lambda,\Lambda}D_{\Lambda,\Lambda}\|_{q\to q} \le 2\,;$$

- $$\|\hat{D}_{\Lambda^c,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda} - D_{\Lambda^c,\Lambda}D^{-1}_{\Lambda,\Lambda}\|_{1,\infty} \le 4s\delta_D.$$

Proof. First, we have

$$\begin{aligned}\|D_{\Lambda,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\|_{q\to q} &= \|I + (D_{\Lambda,\Lambda} - \hat{D}_{\Lambda,\Lambda})\hat{D}^{-1}_{\Lambda,\Lambda}\|_{q\to q}\\ &\le 1 + \|(D_{\Lambda,\Lambda} - \hat{D}_{\Lambda,\Lambda})D^{-1}_{\Lambda,\Lambda}\|_{q\to q}\|D_{\Lambda,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\|_{q\to q}\\ &\le 1 + s\delta_D\|D_{\Lambda,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\|_{q\to q},\end{aligned}$$

which solving the inequality and since $s\delta_D \le 1/2$, turns into

$$\|D_{\Lambda,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\|_{q\to q} \le \frac{1}{1 - s\delta_D} \le 2.$$

Similarly, $\|\hat{D}^{-1}_{\Lambda,\Lambda}D_{\Lambda,\Lambda}\|_{q\to q} \le 2$.

Furthermore,

$$\begin{aligned}\|(\hat{D}_{\Lambda^c,\Lambda} - D_{\Lambda^c,\Lambda})\hat{D}^{-1}_{\Lambda,\Lambda}\|_{1,\infty} &\le \|(\hat{D}_{\Lambda^c,\Lambda} - D_{\Lambda^c,\Lambda})D^{-1}_{\Lambda,\Lambda}\|_{1,\infty}\|D_{\Lambda,\Lambda}\hat{D}^{-1}_{\Lambda,\Lambda}\|_{1\to1}\\ &\le 2s\delta_D.\end{aligned}$$

and

$$\begin{aligned}\|D_{\Lambda^c,\Lambda}(D^{-1}_{\Lambda,\Lambda} - \hat{D}^{-1}_{\Lambda,\Lambda})\|_{1,\infty} \le& \|D_{\Lambda,\Lambda^c}D^{-1}_{\Lambda,\Lambda}\|_{1,\infty}\|\hat{D}^{-1}_{\Lambda,\Lambda}(\hat{D}_{\Lambda,\Lambda} - D_{\Lambda,\Lambda})\|_{1\to1}\\ \le& \|D_{\Lambda,\Lambda^c}D^{-1}_{\Lambda,\Lambda}\|_{1,\infty}\|\hat{D}^{-1}_{\Lambda,\Lambda}D_{\Lambda,\Lambda}\|_{1\to1}\|D^{-1}_{\Lambda,\Lambda}(\hat{D} - D)\|_{1\to1}\\ \le& 2\|D_{\Lambda,\Lambda^c}D^{-1}_{\Lambda,\Lambda}\|_{1,\infty}s\delta_D,\end{aligned}$$

which together give us the second inequality. □

IRTG 1792 Discussion Paper Series 2019

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

001 "Cooling Measures and Housing Wealth: Evidence from Singapore" by Wolfgang Karl Härdle, Rainer Schulz, Taojun Xie, January 2019.

002 "Information Arrival, News Sentiment, Volatilities and Jumps of Intraday Returns" by Ya Qian, Jun Tu, Wolfgang Karl Härdle, January 2019.

003 "Estimating low sampling frequency risk measure by high-frequency data" by Niels Wesselhöfft, Wolfgang K. Härdle, January 2019.

004 "Constrained Kelly portfolios under alpha-stable laws" by Niels Wesselhöfft, Wolfgang K. Härdle, January 2019.

005 "Usage Continuance in Software-as-a-Service" by Elias Baumann, Jana Kern, Stefan Lessmann, February 2019.

006 "Adaptive Nonparametric Community Detection" by Larisa Adamyan, Kirill Efimov, Vladimir Spokoiny, February 2019.

007 "Localizing Multivariate CAViaR" by Yegor Klochkov, Wolfgang K. Härdle, Xiu Xu, March 2019.

008 "Forex Exchange Rate Forecasting Using Deep Recurrent Neural Networks" by Alexander J. Dautel, Wolfgang K. Härdle, Stefan Lessmann, Hsin-Vonn Seow, March 2019.

009 "Dynamic Network Perspective of Cryptocurrencies" by Li Guo, Yubo Tao, Wolfgang K. Härdle, April 2019.

010 "Understanding the Role of Housing in Inequality and Social Mobility" by Yang Tang, Xinwen Ni, April 2019.

011 "The role of medical expenses in the saving decision of elderly: a life cycle model" by Xinwen Ni, April 2019.

012 "Voting for Health Insurance Policy: the U.S. versus Europe" by Xinwen Ni, April 2019.

013 "Inference of Break-Points in High-Dimensional Time Series" by Likai Chen, Weining Wang, Wei Biao Wu, May 2019.

014 "Forecasting in Blockchain-based Local Energy Markets" by Michael Kostmann, Wolfgang K. Härdle, June 2019.

015 "Media-expressed tone, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang K. Härdle, Yanchu Liu, June 2019.

016 "What makes cryptocurrencies special? Investor sentiment and return predictability during the bubble" by Cathy Yi-Hsuan Chen, Roméo Després, Li Guo, Thomas Renault, June 2019.

017 "Portmanteau Test and Simultaneous Inference for Serial Covariances" by Han Xiao, Wei Biao Wu, July 2019.

018 "Phenotypic convergence of cryptocurrencies" by Daniel Traian Pele, Niels Wesselhöfft, Wolfgang K. Härdle, Michalis Kolossiatis, Yannis Yatracos, July 2019.

019 "Modelling Systemic Risk Using Neural Network Quantile Regression" by Georg Keilbar, Weining Wang, July 2019.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2019

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

020 "Rise of the Machines? Intraday High-Frequency Trading Patterns of Cryptocurrencies" by Alla A. Petukhina, Raphael C. G. Reule, Wolfgang Karl Härdle, July 2019.

021 "FRM Financial Risk Meter" by Andrija Mihoci, Michael Althof, Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle, July 2019.

022 "A Machine Learning Approach Towards Startup Success Prediction" by Cemre Ünal, Ioana Ceasu, September 2019.

023 "Can Deep Learning Predict Risky Retail Investors? A Case Study in Financial Risk Behavior Forecasting" by A. Kolesnikova, Y. Yang, S. Lessmann, T. Ma, M.-C. Sung, J.E.V. Johnson, September 2019.

024 "Risk of Bitcoin Market: Volatility, Jumps, and Forecasts" by Junjie Hu, Weiyu Kuo, Wolfgang Karl Härdle, October 2019.

025 "SONIC: SOcial Network with Influencers and Communities" by Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle, Yegor Klochkov, October 2019.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.